Michael Teichmann

A plastic multilayer network of the early visual system inspired by the neocortical circuit

Michael Teichmann

A plastic multilayer network of the early visual system inspired by the neocortical circuit

TECHNISCHE UNIVERSITÄT
CHEMNITZ

Universitätsverlag Chemnitz
2018

Impressum

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Angaben sind im Internet über http://www.dnb.de abrufbar.

Titelgrafik: Michael Teichmann, René Larisch
Satz/Layout: Michael Teichmann

Technische Universität Chemnitz/Universitätsbibliothek
Universitätsverlag Chemnitz
09107 Chemnitz
https://www.tu-chemnitz.de/ub/univerlag

readbox unipress
in der readbox publishing GmbH
Am Hawerkamp 31
48155 Münster
http://unipress.readbox.net

ISBN 978-3-96100-065-4

http://nbn-resolving.de/urn:nbn:de:bsz:ch1-qucosa2-318327

TECHNISCHE UNIVERSITÄT
CHEMNITZ

Professorship Artificial Intelligence

A plastic multilayer network of the early visual system inspired by the neocortical circuit

Dissertation

submitted for the degree of

Doktor der Naturwissenschaften (Dr. rer. nat.)

in the
Department of Computer Science,
Technische Universität Chemnitz

by Dipl.-Inf. Michael Teichmann

Day of the submission: August 13, 2018
Day of the defense: September 14, 2018

Examiner: Prof. Dr. Fred H. Hamker
Frédéric Alexandre Directeur de recherche Inria

Teichmann, Michael
michael.teichmann@informatik.tu-chemnitz.de
A plastic multilayer network of the early visual system inspired by the neocortical circuit
Department of Computer Science, Technische Universität Chemnitz
Chemnitz, September 14, 2018

Abstract

The ability of the visual system for object recognition is remarkable. A better understanding of its processing would lead to better computer vision systems and could improve our understanding of the underlying principles which produce intelligence. In the last decades many models have been published which account for the processing in the visual system. However, few of them covered several cortical areas, used biological plausible plasticity mechanisms, and demonstrated the capability to learn receptive fields comparable to the ones found in the visual cortex. Further, few models can account for the learning of invariant object representations.

We propose a computational model of the visual areas V1 and V2. We implemented the two important layers (4, 2/3) of the feedforward pathway for each area. Within the areas we use a rich connectivity which is inspired by the neocortical circuit. As the first deeper visual cortex model, we combined the three most important cortical plasticity mechanisms. 1) Hebbian synaptic plasticity to learn the synapse strengths of the excitatory and inhibitory neurons, including trace learning to learn invariant representations. 2) Intrinsic plasticity to regulate the neurons response properties and stabilize the learning in deeper network layers. 3) Structural plasticity to modify the connections during network training and to overcome the bias for the learnings from the initial definition of the connections. We trained the network on a continuous stream of natural scenes simulating fixational eye movements. This enables the network to learn invariant representations from the temporal coherence of the input.

We demonstrate the functioning and stability of the proposed plasticity mechanisms. We show that our model neurons learn receptive fields comparable to receptive fields in the respective cortical areas. In line with a new hypothesis, we found that V2 neurons are more sensitive to naturalistic textures than V1 neurons. We also verify the invariant object recognition performance of the model on the COIL-100 dataset. We further show that invariance is build up gradually in the model and that this is impaired without trace learning. We show that the developed weight strengths and connection probabilities de-

pend on the correlations between the neurons, which results from the interplay of synaptic plasticity and structural plasticity. We relate this results to experimental data and confirm the relation found for excitatory connections. We link the findings for the inhibitory connections to the underlying plasticity mechanisms and explain why inhibitory connections often appear unspecific. Further, we demonstrate the efficiency of the learned neuronal code in terms of the sparseness and correlations between the neurons.

With this, we have created a computational model of the early visual system which employs the most important forms of plasticity. The model is more detailed than previous approaches and learns all elements of the complex network in parallel. It can reproduce neuroscientific findings and fulfills the purpose of the visual system, invariant object recognition.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

LTD	Long-term depression
LTP	Long-term potentiation
STDP	Spike-timing dependent plasticity
SVM	Support vector machine
CNN	Convolutional neural network
DNN	Deep neural network
DCNN	Deep convolutional neural network
SNN	Spiking neural network
SDNN	Spiking deep neural network
DoG	Difference of Gaussian
RBF	Radial basis function
VTU	View tuned units
RF	Receptive field
eCRF	Extra classical receptive field
pRF	Population receptive field
RGC	Retinal ganglion cells
LGN	Lateral geniculate nucleus
V1	Visual area 1, primary visual cortex, striate cortex
V2	Visual area 2, secondary visual cortex, extra striate cortex
V4	Visual area 4
IT	Inferotemporal cortex

1 General Introduction

1.1 Motivation

Despite the progress in the field there is a lack of visual cortex models with a sufficient degree of detail and realistic plasticity mechanisms. The most models implement a standard feedforward hierarchy, which can not account for the dynamic properties of the neurons caused by the recurrent processing in the brain. To compensate this shortcoming, the logic of the cortical circuit was often packed into the logic of individual neurons by using nonlinear functions. The simpler architecture goes often in line with simplified plasticity mechanisms or non biological learning principles. In particular the inhibitory circuit and its plasticity is often neglected or the learning of invariant representation is replaced by a simple concept like max-pooling. We believe that a higher degree of detail is required for models with more explanatory power on the underlying cortical processes.

We aim to design a model which is realistic enough to reproduce experimental findings and its mechanisms should also be applicable within more complex models, such as models of visual attention, which can account in detail for the processes within the visual cortex which contribute to the phenomenon of attention. Therefore, we have to implement important parts of the neocortical circuit and we have employed cortical plasticity mechanisms. We find of particular importance in modeling the recurrent interactions over the inhibitory circuit and the ability to learn invariant representations. Finally, the learnings and model responses should be comparable to experimental findings to ensure that the model has the potential to make plausible predictions.

Based on the assumption that computational principles in the cortex are largely similar, we further hope to transfer the gained knowledge to other cortical areas, our model will cover areas which span over about 20 percent of the neocortical surface, and allow the modeling of even larger networks. A better understanding of the underlying principles of cortical processing could lead to better computer vision systems and will improve our understanding how the human brain produces intelligence.

1.2 Thesis structure

The first chapter of the thesis, the General Introduction, introduce fundamental findings from which we derived the architecture and plasticity mechanisms of our model. We go from the general to the detail and explain related neuroscientific findings and assumptions. Subsequently, we develop a concept for our computational model, derived from our knowledge about structure and plasticity of the visual cortex. Finally, we introduce related computational models and give an overview of their structure, plasticity mechanisms, and achievements.

The second chapter describes the concrete architecture of our model. This includes the implemented connectivity on all levels of detail and the amount of neurons and their logical organization. The third chapter explains the used software, stimuli, data preprocessing, and presentation protocol, which we use throughout the thesis.

The fourth to the sixth chapter describe the different plasticity mechanisms of our model. Note, we implement a single computational model, which includes all described mechanisms. However, we split the description into these three chapters. The chapters are ordered by their dependency. First, we introduce the structural plasticity mechanism, which depends on the learnings of the network, but the concrete synaptic plasticity mechanisms are of minor importance. Similarly, the intrinsic plasticity, where we just have to refer to the general concepts of synaptic plasticity. Finally, we introduce the synaptic plasticity and and can profit from the knowledge of the functioning of all the employed plasticity mechanisms in the network. Within each chapter, we first introduce related computational principles and then describe our implementation. Subsequently, we evaluate our model mechanisms. It is intended to give relevant information in a compact form. This means that the relevant neuroscientific foundations and related models are given in the sections describing the implementation. Also the evaluation sections introduce first the questions we address and their importance, followed by a description of the evaluation methods, and the results. We draw conclusions from the results and their relation to other findings immediately to not scatter issues over large parts of the thesis. Each of the chapters has its own conclusion, which gives a comprehensive view on the findings in the chapter.

Finally, we close the thesis with the General Discussion. This section takes up the important findings from the single chapters and relate them to other modeling and experimental studies.

1.3 Visual cortex

The visual cortex is the largest part of the primate brain. It occupies in macaque monkeys about 55 percent of the neocortical surface (or 52 percent of the cerebral cortex) (Felleman and Van Essen, 1991, Table 2). This is much larger than other cortical subdivisions, for instance the somatosensory cortex with 11.5 percent and the auditory cortex with 3.4 percent, or the motor cortex with 7.9 percent (Felleman and Van Essen, 1991, Table 2). The largest areas of the visual cortex are the early visual areas V1 and V2, which cover about 42.9 percent of the visual cortex surface or 23.7 percent of the neocortex (Felleman and Van Essen, 1991, Table 2). Hence, we believe that when we understand the functioning of these large areas, including the mechanisms how the network is formed and the neurons interact, we would be able to model also other parts of the brain with a similar machinery.

The visual system, as well as the whole cerebral cortex, is a highly recurrent structure and not strictly hierarchical, i.e. a large fraction of the areas are interconnected (Felleman and Van Essen, 1991, Tables 3,4). Nevertheless, a forward hierarchy was identified (Felleman and Van Essen, 1991) which form paths from the sensory inputs, the eyes, to higher order areas. Along these paths the neurons of the different areas respond to features with increasing complexity. Two major pathways have been described, the ventral pathway, reaching to the inferior temporal cortex (IT), and the dorsal pathway, reaching to the parietal cortex (Gazzaniga et al., 2009). Along the dorsal pathway properties like movement are processed, also spatial awareness and the guidance of actions. The pathway is called the "where an object is" pathway (Gazzaniga et al., 2009). The ventral pathway is called the "what we're looking at" pathway (Gazzaniga et al., 2009). Because, the covered areas represent properties like the form of an object and facilitate our ability for object recognition. The areas we considered in this thesis are the first two areas (V1 and V2) of the ventral pathway. We do not consider the dorsal pathway, despite its importance and its interconnections to the ventral pathway.

The neurons in the areas along the ventral pathway are characterized by a successive increase in receptive field size and complexity of the features they respond for (Fig. 1.1). This is accompanied with an increasing invariance against the exact appearance of their preferred stimuli (DiCarlo et al., 2012). Invariance means the ability to recognize stimuli, like objects, independent from their position, rotation, or scale. We evaluate this aspect for our model in Section 6.3.4. It is believed that invariance is gradually increased over the hierarchy (DiCarlo et al., 2012). This has the advantage that the neurons keep their sen-

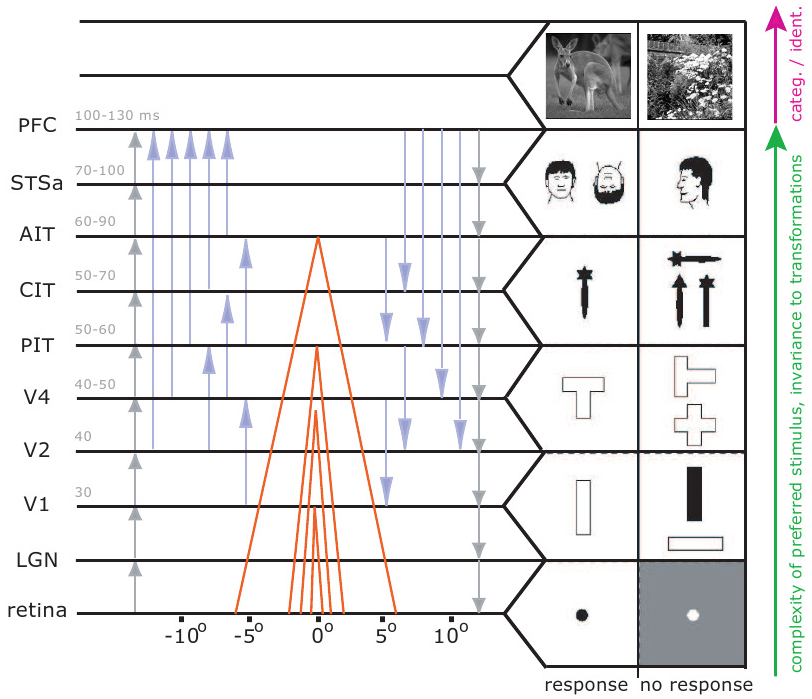

FIGURE 1.1: "The organization of visual cortex based on a core of knowledge." Taken from Serre (2006, Fig. 1-2). Left, the visual areas of the ventral stream are listed by their assumed position in the hierarchy. The response delays from stimulus onset are given in gray. The main feedforward pathway is also given by the gray upward arrows. The gray downward arrows denote the main feedback pathway. The blue arrows illustrate forward and backward short-paths between the areas. In the middle the orange lines indicate the receptive field sizes, given in degree on the x-axis. On the right, example stimuli are illustrated for which the neurons in the different areas are assumed to respond and besides stimuli where the same neurons would not respond are shown.

sitivity to the relative position of the features of their preferred stimulus (Földiák, 1998). It is hypothesized, and logical, that the areas along the hierarchy improve the linear separability of the input (DiCarlo and Cox, 2007). This means that objects, that are difficult to separate in the image space (Fig. 1.2a), become more and more "untangled", i.e. linear separable, in the neural space (Fig. 1.2b). We evaluate this aspect for our model in Section 6.3.5.

1.3.1 Primary visual cortex (V1)

The primary visual cortex is the most investigated area of the visual cortex. It receives input from the cells in the lateral geniculate nucleus (LGN), a thalamic area, which by itself receives inputs from the retinal ganglion cells (RGC) in the eyes. LGN neurons

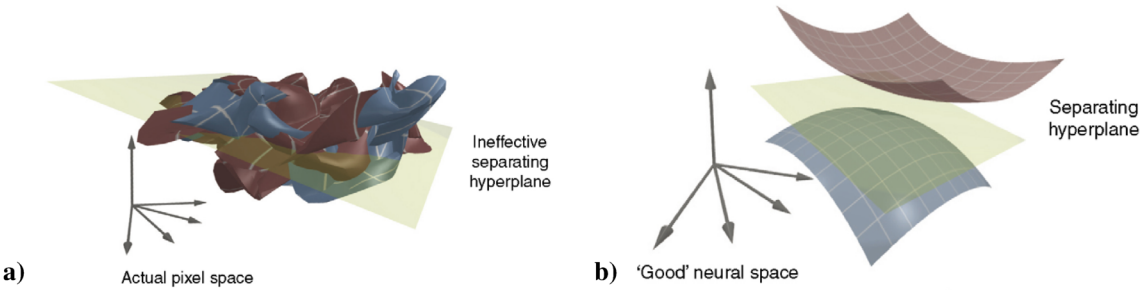

FIGURE 1.2: "Illustration of object tangling." Taken from DiCarlo and Cox (2007, Fig. 1). **a)** Two different images (indicated by blue and red) in the input space. The images are not separable through a simple hyperplane, they are tangled. **b)** The same images in a good neuronal space. Both images are linear separable, they are untangled.

respond similar to the RGCs. Their receptive fields are typically described as circular fields with a small center and a larger surround, where the center is whether excitatory (on) or inhibitory (off) and the surround has the opposite property (Carandini et al., 2005; Gazzaniga et al., 2009). Thus, they respond best to spots of light (Hubel and Wiesel, 1962). As typical in the brain the retina and LGN are much more complex than they are treated in computational models (Carandini et al., 2005). We regard for our model only the parvocellular cells of LGN, which are visually driven over "red" and "green" cones. These cells transport information over color and form. Other cell types in LGN are the magnocellular cells (driven by rods) and koniocellular cells ("blue" cones). The LGN inputs to V1 mainly target the middle layer, layer-4, which in turn projects to neurons in layer-2/3 (Douglas and Martin, 2004). The neurons in layer-2/3 again project to deeper areas as the area V2. A more detailed description of the V1 circuit is given in Section 1.4.

The behavior of the neurons in V1 is typically described as they connect the outputs of their preceding layer to form a more complex receptive field (Hubel and Wiesel, 1962). In layer-4, the entrance layer of V1, primarily neurons are found responding to bar like patterns. The receptive fields of these neurons could be easily described through maps of regions where light has an excitatory and inhibitory effect on the neuron activity. According to this, the neurons are called "simple-cells" (Hubel and Wiesel, 1962). It was found that Gabor-wavelets are a good mathematical description for these receptive fields (Jones and Palmer, 1987). The neurons in the next stage, layer-2/3, are not that easy to describe. It was found that the most neurons here respond to stimuli with similar properties as the simple-cells respond for, however, a map of positions where light has an excitatory or in-

hibitory effect could not be made (Hubel and Wiesel, 1962). This is probably because of the invariance of the neurons to the exact position of a stimulus which leads to overlapping subfields (Schiller et al., 1976; De Valois et al., 1982; Carandini et al., 2005; Martinez et al., 2005). Consequently, the neurons have been named "complex-cells". For a more detailed description of the receptive field properties please see Section 6.3.2. For a review see Carandini et al. (2005). In their groundbreaking work Hubel and Wiesel described the mentioned receptive field properties (Hubel and Wiesel, 1962, 1968). Further, they introduced the simple conceptual scheme for the hierarchy of simple and complex-cells (Hubel and Wiesel, 1962, Fig. 20). This scheme of two consecutive stages, where the first stage combines the inputs to new features and the second stage connects the neurons responding to similar features regardless of their exact position, inspired a large number of computational models. One of the first models was the Neocognitron, which stacked three stages of "simple" and "complex" layers over each other and was used for hand-written digit recognition (Fukushima, 1980). A model resembling the properties of simple and complex-cell responses was the Gabor energy model of Adelson and Bergen (1985), it used Gabor functions to mimic simple-cells and squared these functions, similar to add a (phase) shifted Gabor, to account for complex-cell responses. Another prominent example is the HMAX model, which introduced a maximum function as function for the "complex" layers (Riesenhuber and Poggio, 1999; Serre et al., 2007). Similar mechanisms are used in the nowadays very prominent model class of deep convolutional neural networks (CNN). One of the first approaches is the LeNet-5, it consists of three convolutional layers, where the first two layers are followed by subsampling layers which pool neighboring similar features, the last convolutional layer was followed by two fully connected layers (LeCun et al., 1998). In newer CNNs the subsampling was replaced by a maximum operation (max-pooling) (e.g. Szegedy et al., 2014). Moreover, the development of simple-cell receptive fields became a testbed for models of synaptic plasticity (e.g. Falconbridge et al., 2006; Wiltschut and Hamker, 2009; Clopath et al., 2010; King et al., 2013). A proper model of the plasticity mechanisms should be able to learn Gabor like receptive fields of different orientations and spatial frequencies.

1.3.2 Secondary visual cortex (V2)

Area V2 was less often investigated. This might have its reason in the unclear behavior of its neurons. The receptive fields are more complex than in V1 and difficult to catego-

rize. The natural hypothesis would be that the neurons respond best to contours or angles (corners) the next more complex types of stimuli in comparison to V1. However, a large fraction of neurons respond to similar stimuli as V1 neurons and with similar intensity on more complex stimuli as contours or textures (Kobatake and Tanaka, 1994; Hegdé and Van Essen, 2000). A difference between V1 and V2 was found in their responses to naturalistic textures. V2 neurons showed a higher response by average to naturalistic textures than V1 neurons, in comparison to noise images with the same spatial frequency structure (Freeman et al., 2013). For a more detailed description of the receptive field properties please see Section 6.3.2.

The connections and layered structure of V2 is assumed to be similar to V1 (Douglas and Martin, 2004) and will be introduced on the example of the V1 circuit in the next section. Despite, the inner structure of the areas is similar, a differentiation of receptive field types in simple and complex-cells has not been made in V2. When following the definition of complex receptive fields anyway all receptive fields in deeper layers than V1 layer-2/3 are presumably as complex. Nevertheless, computational models assume the same stacking of operations (feature extraction and pooling) in the areas succeeding V1 (examples are given in the previous section). V2 layer-2/3 by itself projects to area V4, the next area in the ventral pathway, which we do not regard in this thesis.

1.4 Circuit of a neocortical area

As described above, visual areas share a basic hierarchy of processing. This hierarchy seems to be present through out the neocortex (Douglas and Martin, 2004). However, physiological data largely differ in amount, quality, and detail for the different areas. Whereas the circuit within V1 is well studied, the circuit of V2 is poorly examined (Sincich and Horton, 2005). The inter area connections are better underpinned with data.

Without knowing much about the connectivity in the primary visual cortex, Hubel and Wiesel (1962) derived a basic architecture from the properties of the receptive fields they found in the layers 4 and 2/3, where the complex-cells are found. This is that LGN projects to layer-4, where the simple-cells are found, and layer-4 projects to layer-2/3. However, a neocortical area consists typically of six anatomical layers. Where the internal layer (layer 4) receives inputs from the preceding area, or the thalamus, and projects to the superficial layers 2 and 3, which in turn project to the internal layer of the next area (cortico-cortical)

(Thomson and Bannister, 2003; Douglas and Martin, 2004; Sincich and Horton, 2005; Shipp, 2007; Anderson and Martin, 2009). Both, internal and superficial layers project to the deep layers, layer-5 and 6, which serve the inter area communication over the second order thalamus, the pulvinar (cortico-thalamic-cortical) (Shipp, 2003, 2007; Thomson, 2010; Shipp, 2015). Further, layer-6 sends feedback to the preceding area, to layer-6 for cortical areas (cortico-cortical) (Sincich and Horton, 2005; Shipp, 2007; Thomson, 2010) or, in the case of V1, to LGN (cortico-thalamic) (Shipp, 2003, 2007; Thomson, 2010; Shipp, 2015). Layer-1 is an exception in this structure. It contains nearly no cell bodies, however, the feedback axons from the pulvinar as well as from layer-2/3 and layer-6 neurons of the succeeding area terminate there (Rockland and Virga, 1989; Shipp, 2007; Anderson and Martin, 2009). These axons connect to dendrites, located in layer-1, of layer-2/3 and layer-5 neurons (Shipp, 2007).

A more detailed concept on the wiring within area V1 was developed by Potjans and Diesmann (2014). They combined the anatomical data about connections between the two principle neuron types, excitatory and inhibitory neurons, in the different layers and formed a circuit description suitable for computational modeling. This circuit model contains the connection probabilities between the two principle neuron types in the different layers(Potjans and Diesmann, 2014, Table 5). In general all neighboring neuron populations share connections. However, there is some specificity (Fig. 1.3). Inhibitory neurons are mainly locally connected to excitatory and other inhibitory neurons in the same layer. Excitatory neurons are locally connected to other excitatory neurons and the inhibitory neurons in the same layer. Between the layers, the connection probabilities draw the same two major pathways of information flow between excitatory neurons as described above based on the qualitative anatomical data. One pathway from layer-4 to layer-2/3 and a second from layer-4 and layer-2/3 to layer-5 and from there to layer-6. Layer-5 again modulates the activity of layer-2/3 by targeting a larger quantity of its inhibitory neurons. Whereas layer-6 modulates layer-4, also by targeting more inhibitory neurons.

Because of the lack of detailed information, we assume for this thesis that the internal structure in V2 shares large similarities to V1. For the computational model we introduce we will not regard the layers 5 and 6. This is because they are strongly involved in inter area communication over the thalamus and do not project directly to area V2. So that we would have to model also the second order thalamus and the full feedback circuit between the areas to make use of these layers. To our knowledge only a version of the Leabra

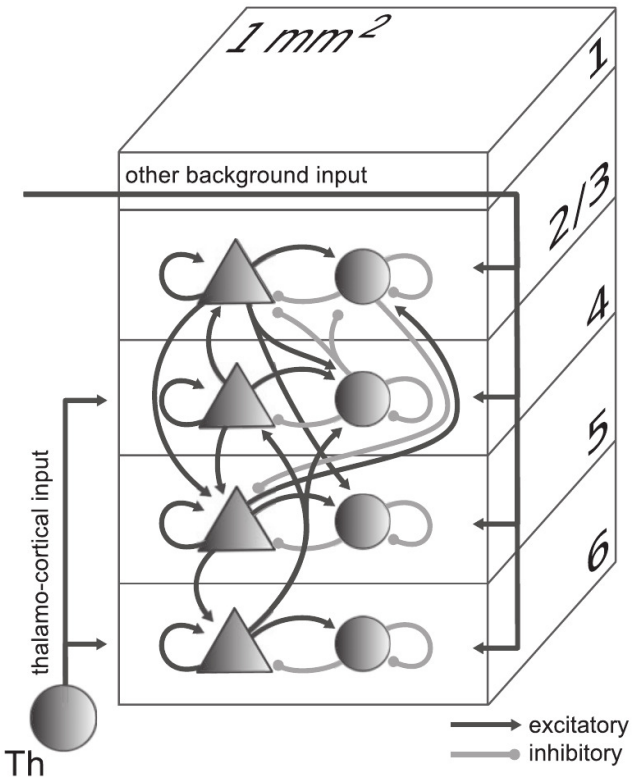

FIGURE 1.3: "Model definition. Layers 2/3, 4, 5, and 6 are each represented by an excitatory (triangles) and an inhibitory (circles) population of model neurons. Input to the populations is represented by thalamo-cortical input targeting layers 4 and 6 and other external background input to all populations. Excitatory (black) and inhibitory (gray) connections with connection probabilities >0.04 are shown." Taken from Potjans and Diesmann (2014, Fig. 1).

framework implements a more complete architecture of the inner and outer connections between neocortical areas, namely LeabraTI (O'Reilly et al., 2014; Kachergis et al., 2014) and DeepLeabra[1]. The model takes the same assumption as we do that V2 is internally functioning similar to V1. However, it focuses on the temporal aspect of the processing for sequential actions (Kachergis et al., 2014) and the influence of temporal context (O'Reilly et al., 2014). Further, the internal learning mechanisms differ from ours. Most importantly in the use of a k-winner inhibitory mechanism and a (prediction) error-driven learning rule.

[1] `https://grey.colorado.edu/CompCogNeuro/index.php/DeepLeabra`

1.5 Neurons

As mentioned before, there are two principle neuron types in the cortex: excitatory and inhibitory neurons. There are much more methodologies to differentiate into several subtypes, e.g. morphology, synaptic transmitters, firing pattern. To reduce the model complexity we will focus on the two principle types. Further, as described in the previous section, data for the design of the network circuit are available on this detail level.

1.5.1 Excitatory and inhibitory neurons

Neurons can be grouped in two main classes by the effect they have on other neurons. Excitatory neurons drive other neurons, i.e. their activity increases the membrane potential of the postsynaptic neuron. Whereas inhibitory neurons inhibit other neurons, i.e. their activity decreases the membrane potential. Inhibitory neurons are also often called interneurons as they are often just locally connected (cf. previous section). The membrane potential has to exceed a certain threshold to make the neuron active to effect other neurons. The synapses of the neurons have always the same chemical properties (Dale's law). Thus, an excitatory neuron will form only synapses having an excitatory effect, similar inhibitory neurons will have just inhibitory synapses. Hence, the synapse type determines the type of the neuron in that kind of classification.

Early computational models rarely implemented inhibitory interneurons (e.g. Földiák, 1990, 1998; Riesenhuber and Poggio, 1999; Falconbridge et al., 2006; Serre et al., 2007; Masquelier and Thorpe, 2007; Wiltschut and Hamker, 2009; Savin et al., 2010; Zylberberg et al., 2011; Teichmann et al., 2012; Masquelier, 2012; Willmore et al., 2012). Whether they implemented direct lateral inhibitory connections between the otherwise excitatory neurons (e.g. Földiák, 1990, 1998; Falconbridge et al., 2006; Wiltschut and Hamker, 2009; Savin et al., 2010; Zylberberg et al., 2011; Teichmann et al., 2012; Masquelier, 2012), or they used different plasticity schemes which do not require explicit inhibitory connections (e.g. Serre et al., 2007; Masquelier and Thorpe, 2007; Willmore et al., 2012). In recent years more models implemented also inhibitory interneurons (e.g. Vogels et al., 2011; King et al., 2013; Diehl and Cook, 2015; Sadeh et al., 2015; Miconi et al., 2016). However, in some networks the plasticity is restricted to the inhibitory connections only (e.g. Vogels et al., 2011) or to the excitatory connections only (e.g. Miconi et al., 2016). Others have non plastic feedforward connections or fixed direct inputs to the neurons, strongly

determining the correlations between the neurons (Sadeh et al., 2015).

1.5.2 Spike vs. rate

Despite it is a long time known that neurons communicate via spikes many models use a continuous variable representing the firing rate of a neuron instead of simulating single spike events (e.g. Földiák, 1990, 1998; Riesenhuber and Poggio, 1999; Falconbridge et al., 2006; Serre et al., 2007; Wiltschut and Hamker, 2009; Teichmann et al., 2012; Willmore et al., 2012). The firing rate can be interpreted as the amount of spikes within a certain time interval. For instance a firing rate of 100 can stand for a spike frequency of 100Hz. However, often the rate is taken as an arbitrary number, where inactivity is symbolized by zero and positive values denote activity on an arbitrary scale. A spike representation has the advantage that the causal effect of a single spike can be taken into account for the synaptic learning. Hence, for us causes the choice of the learning rule the need of whether spike or rate representations on the neuron side. This means, when we would use a rule out of the spike-timing dependent plasticity (STDP) class of learning rules, we would require spiking neurons, whereas with other learning rules a rate based representation is natural as no conversion from spike to rate is needed. Thus, the important question to ask is whether the exact spike timing, exploited in STDP learning, is important for the emergence of the neuron properties in the visual cortex or is the learning dominated by the spike frequency, i.e. rate (cf. Brette, 2015). In a study within the rat visual cortex, with systematical variation of the rate and spike timing, it was found that frequency dominates the effect of the timing (Sjöström et al., 2001). Low postsynaptic firing induced LTD and high frequencies induced LTP, independent from the time difference of the pre- and postsynaptic spike. Thus, we preferred rate coded neurons for our model implementation.

1.5.3 Activation function

The activation function of a neuron transfers the input the neuron receives into its output activity. We will focus here on activation functions for rate neurons in computational models, but the assertions we take are not limited to them. Sigmoidal activation functions have been often used to simulate biological neurons. Beside a large approximately linear range, their properties can account for spontaneous activities evoked by weak stimulations and a saturation effect evoked by strong stimulations. Further, a sigmoid function can be con-

figured to account for the shape of the contrast-response function observed in the visual cortex (Albrecht and Hamilton, 1982). However, experimental studies raised doubts on the saturation effect as part of the activation function. In macaque monkeys it was found that the neurons "amplified linearly and do not reach saturation" (Ringach and Malone, 2007). Further, with pharmacological blocking of inhibition it was observed that the maximum response increased multiple times and reaches values of several 100Hz (Katzner et al., 2011). Also, it seems that the presentation time has an effect on the saturation, it was found that short presentation times lead to more linear contrast-response functions than long one (Dai and Wang, 2018), which indicates saturation as a network effect. Typically contrast normalization is explained through divisive normalization, which is again explained as the result of the inhibitory circuit (Heeger, 1992; Kouh and Poggio, 2008; Graham, 2011). In the deep learning community rectified linear units (ReLU) became widely used in recent years. This is because they are faster to compute and do not saturate, which would impair the network convergence (Krizhevsky et al., 2012). Rectified linear unit means the output is zero when the neurons activity would be below a certain threshold and increases linearly for higher values. This matches largely the function derived from experimental observations by Ringach and Malone (2007), who found a rectification followed by a small nonlinear transition zone and finally a linear function. The transition zone had been explained as effect of Gaussian noise in the input.

1.6 Forms of plasticity

In the brain several forms of plasticity cooperate with each other. The most prominent form of plasticity is the synaptic plasticity (Sec. 1.6.1). This is how the connections between the neurons, the synapses, change their efficiency. However, the brain has billions of such synapses which are physically independent from each other. Also the synapses reaching a single neuron are just indirectly connected, e.g. over the backpropagated action potential of the neuron. Despite this independence the synaptic weights have to change in a controlled fashion so that the neurons receive inputs in a meaningful range.

Two presumably different mechanisms have been found to stabilize the neurons functioning. Synaptic homeostasis and intrinsic homeostasis (Turrigiano and Nelson, 2004; Turrigiano, 2011). Homeostasis means here the process of self stabilizing the neurons activity. Synaptic homeostasis is often treated as part of the synaptic plasticity. That is, the

change of the weight strength underlies a kind of normalization process, holding the neurons activity in a certain range. Intrinsic homeostasis also stabilizes the neurons activity, but independent from the synaptic weight strengths (Zhang and Linden, 2003). Thus, it is a form of plasticity independent from the synaptic plasticity and therefore named intrinsic plasticity (Sec. 1.6.2). Intrinsic, because of the change of the neuron's internal electric properties, i.e. how strong it responds on a certain input current.

Synaptic learning is also accompanied by structural changes. Two major processes take place in the brain: neurogenesis and changes in the synaptic structure (Sec. 1.6.3) as neurite outgrowth, synapse formation, and synaptic rewiring (Butz et al., 2009b). Neurogenesis plays a minor role in the visual cortex, thus we do not regard it in this thesis. Whereas retraction and growth of dendrites or axons and the removal or formation of synapses are taking place the whole life (Holtmaat and Svoboda, 2009; Caroni et al., 2012). When we refer to structural plasticity in this thesis we mean these processes. Specifically, we focus on the formation and removal of synapses, as we have no explicit model of the dendrite or axon implemented and thus can just indirectly account for such changes.

1.6.1 Synaptic plasticity

The development of the visual system is experience dependent (Ohzawa and Freeman, 1988). This means, the visual system learns its properties from the responses evoked by the light falling on the retina. The learning occurs by changes in the synaptic efficiencies of the efferent neurons. Note, we will use in this thesis the terms learning and synaptic plasticity synonym. Further, we will refer to the synaptic efficiency with the more implementation related term weight strength, or simply weight. Two major processes have been described for the efficiency change: long term potentiation (LTP) and long term depression (LTD) (Malenka and Bear, 2004; Feldman, 2009). LTP describes the increase of synaptic efficiency and LTD the decrease. LTP can be induced through a short strong stimulation, a longer weak stimulation induces LTD. For a review see Malenka and Bear (2004) or Feldman (2009). Among this rate sensitivity, a timing dependent process has been discovered (Bi and Poo, 1998; Caporale and Dan, 2008). This is called spike-timing dependent plasticity (STDP). Beside differences between synapse type and cortical region, LTP is induced by a presynaptic spike few milliseconds followed by a postsynaptic spike. One could say the causality of both events triggers the increase of the weight. LTD is induced by the contrary event, i.e. when a postsynaptic spike is followed by a presynaptic spike.

However, under natural conditions this is much more complicated. LTP and LTD inducing spike pairs can occur in short time after each other. It was found that the weight change depends on both: timing and firing rate (Sjöström et al., 2001; Caporale and Dan, 2008). For a review see Caporale and Dan (2008). Among all details, the weight change follows the Hebbian paradigm (Hebb, 1949). This means it is an associative process which bases on the coincidence of the presynaptic and postsynaptic activity. Therefore the plasticity on excitatory synapses is called Hebbian (Dayan and Abbott, 2001, Sec. 8.3). The plasticity on the inhibitory synapses is called anti-Hebbian because of the different, inhibiting, effect these synapses have on the postsynaptic activity.

Since the 80's an increasing amount of research articles has been published about synaptic plasticity. It is a vital field with more than thousand articles per year for more than a decade (Nelson and Turrigiano, 2008, the values remain on this level till today). Despite this large corpus of literature no concrete mathematical model of synaptic plasticity was established. Because of that, we focus on a class of learning mechanisms which have shown to be able to learn receptive fields similar to the one of V1 simple-cells. Convincing sets of simple-cell receptive field have been learned by rate based learning rules combining a Hebbian term with Oja normalization (Oja, 1982) and anti-Hebbian learning for the decorrelation of the neurons (Falconbridge et al., 2006; Wiltschut and Hamker, 2009). In both models the receptive fields are Gabor like (Jones and Palmer, 1987). Wiltschut and Hamker (2009) could also show a good similarity to the properties of Gabor fits from macaque monkey neurons (cf. Ringach, 2002). Similar mechanisms have been successfully used in network models using spiking neurons (Zylberberg et al., 2011; King et al., 2013). In contrast to these mechanisms, some STDP rules fail in developing proper simple-cell receptive fields (e.g. Clopath et al., 2010). That is because their normalization scheme leads to U-shaped weight distributions (Morrison et al., 2008), i.e. all weights cluster close to the maximum or minimum weight value (additive STDP). This does not resemble a proper Gabor function with gradual decreasing weights. For an introduction in related rate-based synaptic plasticity rules please see Section 6.1.

STDP rules as well as rate-based synaptic plasticity rules underly the same limitation. Both need a mechanism to restrict the infinite growth of the synaptic weights (normalization). That is because of the Hebbian character of the rules. When coactivity leads to an increase of the weight it becomes causality and the weight grows unlimited. Thus, the weights have to be bound, a homeostatic process is needed. In experimental stud-

ies a homeostatic process called synaptic scaling has been found (Turrigiano and Nelson, 2004; Feldman, 2009; Turrigiano, 2011). This process can change the receptor numbers of synapses globally to stabilize the activities (Turrigiano and Nelson, 2004; Feldman, 2009; Turrigiano, 2011). Turrigiano et al. (1998) showed that when blocking activities the amplitudes of the postsynaptic currents increased, when measured, while the relative strengths remained stable. When the inhibition was blocked, i.e. the activity was increased, the measured amplitudes slowly decreased and the activities reached the control level. Also Desai et al. (2002) found synaptic scaling in rats. They observed a continuous decrease of the amplitudes with development, which was reduced by dark rearing. Further, monocular deprivation up-scaled the measured amplitudes. They again report this scaling as a global effect. For a review see Turrigiano (2011). For computational principles of synaptic scaling please see Section 6.1 as well.

1.6.2 Intrinsic plasticity

The long-lasting change of the electrical properties of a neuron is called intrinsic plasticity. It is typically caused by changes in the neurons activity. Electrical properties meant its intrinsic excitability. "Intrinsic excitability is the electrical excitability of a particular neuron. It is determined by the number and distribution of ion channels and receptors that contribute the electrical properties and depolarization potential of the neuron."[2]. These changes are a non-Hebbian form of plasticity which stabilizes the firing of a neuron. Desai et al. (1999) discovered that long-lasting changes of neuron activity effect their intrinsic electrical properties. They measured a strong increased intrinsic excitability after two days of deprived activity, in terms of increased firing rates for the same stimulation protocol. Moreover, the neurons respond also to weaker currents, i.e. the spike threshold decreased. To control whether synaptic or intrinsic changes caused the observations they pharmacologically blocked synaptic transmissions. They found that activities regulate the ionic conductances. Intrinsic plasticity is intended to serve a mechanism for homeostatic regulation of the neuronal activity to preserve the operating point of the neurons (Turrigiano, 2011). It was hypothesized that the goal of the brain is to maximize the entropy under the restriction of the metabolic costs (Triesch, 2005b). Maximal entropy can be achieved when a single neuron with fixed mean (metabolic costs) approaches an exponential firing distribution (Levy and Baxter, 1996; Triesch, 2005b). This idea inspired early models of

[2] https://www.nature.com/subjects/intrinsic-excitability

intrinsic plasticity. For computational principles of intrinsic plasticity please see Section 5.1.

1.6.3 Structural plasticity

It is well known that experience leads to changes in the synaptic strengths (synaptic plasticity). In the developing brain this is accompanied with large reorganizations of the connectivity (Holtmaat and Svoboda, 2009). Regarding the receptive field of a single neuron, the effect of visual experience on the receptive field structure in the developing visual system has been demonstrated by Vislay-Meltzer et al. (2006). In their study they exploit that a pattern is learned when a postsynaptic spike is released shortly after stimulating the neuron (STDP). While presenting various stimuli at the "target" and "non-target" location beside the receptive field location, they controlled the postsynaptic activity (Fig. 1.4). This means, for stimuli at the target location a spike is manually evoked in the postsynaptic neuron shortly after presentation. It has been observed, using rapid reverse correlation mapping, that the receptive field center is moving to the area of this paired stimulation, demonstrating the effect of visual experience on the synaptic weights. It can be assumed that this process is accompanied with structural plasticity in the developing as well as in the adult brain (Caroni et al., 2012). This implies that new synapses are formed at the target location and older not anymore used synapses are eliminated. Hence, structural plasticity is a process accompanying other plasticity processes.

Cellular manifestations are the central mechanism underlying memory formation (Kasai et al., 2010). Further, structural plasticity gives neurons a tool at hand to maintain their activity level beyond the limitations of synaptic or intrinsic modifications. That is, neurons can overcome insufficient or pathological driving inputs by not relying on them and forming new synapses or eliminating unwanted (Butz et al., 2009b). Thus, structural plasticity increases the robustness of the brain to malfunctions, but also facilitate learning beyond the innate neural connectivity. Moreover, the brain can maintain, with a limited amount of synapses, a wide range of functioning. Without changes in the synaptic connections each neuron would need a tailored connection structure for its individual functioning, which has to be optimal for all phases of the development from the formation of the nervous system to adulthood.

The cortical tissue is very dense, many dendrites and axons are within a small area.

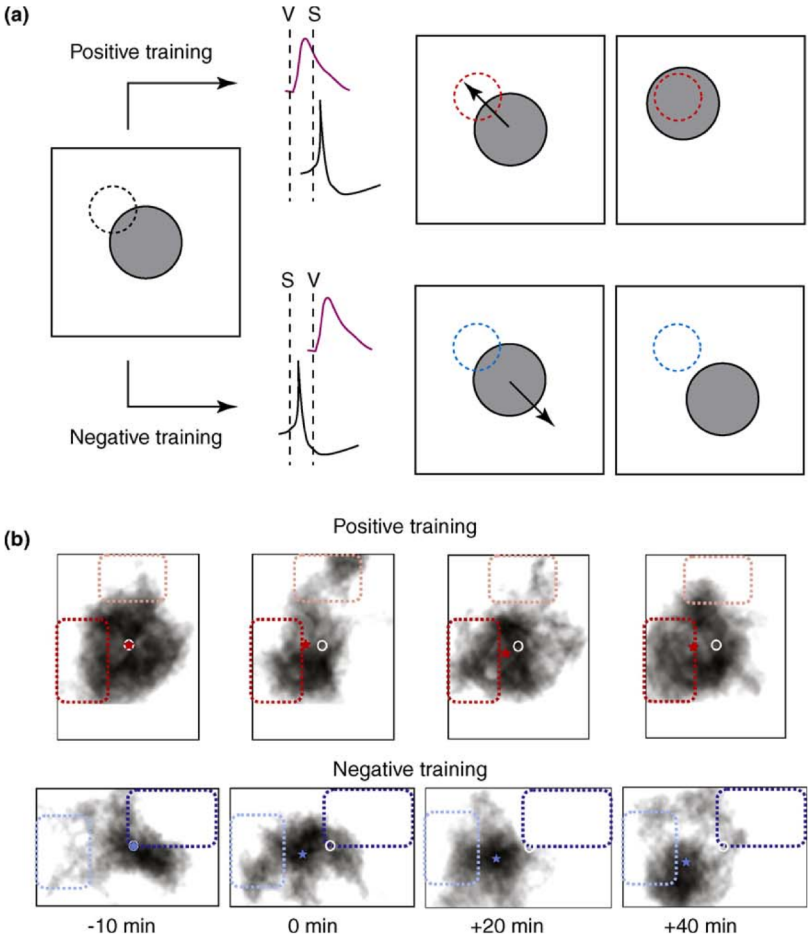

FIGURE 1.4: Effect of visual activity on the spatial receptive field properties. Induced activity while controlling the postsynaptic spike timing can alter the spatial organization of receptive fields. For positive training postsynaptic spikes are evoked at timings inducing LTP, for negative training they are evoked to induce LTD. **(a)** Scheme of the training protocol. Positive training moves the receptive field (gray area) center to the stimulus location (dashed circle) and negative training has the opposite effect. **(b)** Data from tectal cells in Xenopus tadpoles. Dark dotted boxes indicates the stimuli positions. Brighter dotted boxes are the unpaired control region. The circle indicates the original receptive field center, the star indicates the new center. Taken from (Ruthazer and Aizenman, 2010).

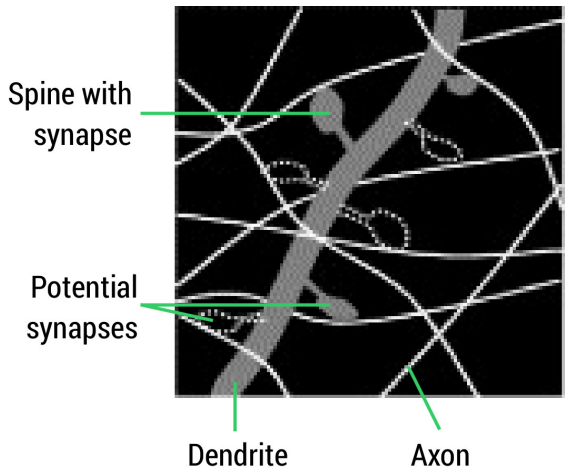

FIGURE 1.5: Illustration of synaptic connectivity patterns. Spines can spontaneous grow out from a dentridic arbor (gray) to neighboring axons and form new synapses. The set of positions where a dendrite is close enough to axons to form synapses is called potential synapses. Modified from (Stepanyants and Chklovskii, 2005).

Thus, dendritic spines[3] can bridge the distance to neighboring axons and easily form new contacts (Fig. 1.5). The set of positions where a dendrite is close enough to axons to form synapses is called potential synapses (Stepanyants et al., 2002). It has been found that it is likely that new synapses are formed in the vicinity of existing strong synapses (Caroni et al., 2012; Harvey et al., 2008; De Roo et al., 2008). These newly formed synapses are originated by thin spines (Knott et al., 2006). The spine volume is related to the synapse strength (Matsuzaki et al., 2004; Holtmaat and Svoboda, 2009). That is, thin spines have weak synapses. Furthermore, the lifespan of a spine is also correlated to its volume. Thin spines are likely to disappear soon (Yasumatsu et al., 2008), while spines with larger volume are found being more stable (Knott et al., 2006; Yasumatsu et al., 2008). Newly formed synapses are subjected to normal synaptic plasticity promoting stabilization (Caroni et al., 2012; Holtmaat et al., 2005). Structural plasticity mechanisms differ between neuron types. We assume for this thesis similar core principles for all components of the computational model. For related computational principles of structural plasticity please see Section 4.1.

1.7 Developing a model sketch

In the previous sections we introduced the fundamental findings we need for developing a computational model. Our goal is to model the underlying circuit of neocortical pro-

[3]"Small protrusions of the dendrite with which an axon terminal forms a synapse." (Spine (Dendritic) Binder et al., 2009)

cessing. Universal enough to be applicable to other brain areas. Precise enough to have explanatory power on the processes in the early visual system.

Basic elements The core units of the intended network model are the excitatory neurons. This is because they are the only units who project to other brain areas and gather and transform the information of preceding areas. To the excitatory neurons, we need a form of inhibition to be able to implement any form of biologically plausible synaptic plasticity. Consequently, we will model inhibitory interneurons as dedicated neuron type.

Circuit While the inhibitory interneurons have to be mainly locally connected, the excitatory neurons can have a richer connectivity. The minimal required connectivity would be the feedforward pathway across the brain. That is, a connection from a population of excitatory neurons from the entrance layer of an area (layer 4) to the next downstream layer (layer 2/3) and from the excitatory neurons there to the next area. This circuit seems over idealized. So that a model aiming to allow new insights in the functioning of the brain should implement also the other layers and the recurrent connections between them and between the areas. On the other side, this would potentiate the complexity of the model. Implementing the layers 5 and 6 would make just sense when the cortical feedback circuit including the second order thalamus and its processing are implemented. It seems not promising to start with the complete complexity at once. Thus, we should implement the layers of the feedforward pathway, but include the recurrent connections between the neurons. Therefore, we use as inspiration the findings about the intra connectivity of a typically neocortical area.

Stimuli To measure later neuron responses which are comparable to experimental data we need to stimulate the network with realistic input. The logical consequence from that is using natural scene input for a model of the visual system. This input has to be given as presynaptic activity onto the V1 layer-4 neurons. Of course, in the visual system a cascade of retinal and thalamic processing has taken place until this stage. Albeit, the receptive fields of LGN neurons follow a simple description, which is easy to approximate. This approximation appears standard in computational modeling, thus, we can focus on the cortical processing and use established image preprocessing methods. A neuron layer, which should obviously be named LGN, should transfer the image values into neuron activities.

Plasticity Now we have a good idea in mind how the processing elements of the network should be arranged. But we still have to define the synaptic weights and the neuron functioning. The receptive fields of neurons in the entrance layer are described as Gabor like. Thus, we could define the weights through a mathematical function. However, finding a good parametrization for high biological plausibility would be complicated. Even basic properties as receptive field sizes differ between experimental studies and measures (cf. eCRF and pRF Angelucci et al., 2002; Wandell and Winawer, 2015). Moreover, no concrete data are available for inhibitory connections or the recurrent connections, apart from the feedforward pathway, or the connections within deeper layers, which are even more abstract than their receptive field, which is defined in the input space. Thus, hard wiring the neurons would be tough.

A much more appealing strategy is to implement the self-organizing mechanisms of the cortex itself. This means at first, the synaptic plasticity. We will use rate based learning rules, since they have shown to develop proper V1-like receptive fields. Because of the unbound increase of the weights in naive Hebbian rules, we have to combine the learning with a normalization term which applies a multiplicative normalization without violating the Hebbian property of locality. The multiplicative normalization has also been shown to give proper receptive fields shapes. Further, the rule should be able to account for the invariance properties of neurons, like the one in V1 layer-2/3. The plasticity mechanisms determine the neuron model so that we use the simplest, but plausible enough, activation function for the neurons: a rectified linear function. Albeit synaptic plasticity finds the connection strength between the neurons for us, we would have to connect each individual neuron of two connected neuron populations. This is an ill concept considering about 190 million neurons just in the macaque V1. Again, we could define a connection matrix, individual for each neuron, based on a mathematical concept considering the findings on the receptive field size. This network design would have the potential to give us sufficient insights in the brain machinery, but it will be again strongly biased from our modeling decisions, how we fill up the unknown parameters. To overcome the “modeler’s bias”, we will apply structural plasticity to refine an initially defined connectivity. Structural plasticity should be a random process just weight based. To not degenerate to a fully random process, dealing with the enormous amount of potential connections which could exist in the brain, the physiological property that new synapses are just formed where dendrites are should be used. Not required synapses should be removed based on their

weight strength. Luckily, this has also support by experimental studies. The network has yet no constraint which restricts the neurons to encode all important information. Any neuron has just access to local information, no global signals, as image reconstruction errors, are used in this network concept. What we realized after a plenty of unsuccessful model runs, should be given here as the third required plasticity principle. We need a mechanism to stabilize the operating point of the neurons, which enforces a somehow equal distribution of information across the neuronal code, i.e. intrinsic plasticity.

All together these principles should enable a modeler to build a functioning, rich, and appealing model of the early visual system. There are other possible design decisions, which might lead to models with higher explanatory power, however, we create our work on certain goals and hypothesis.

Underlying assumptions The first and most important one is that experience shapes the brain. This means that the morphology of the neurons, the structure of the connections, the intrinsic and response properties are assumed to be formed by modifications through the plasticity mechanisms in response to experience (sensory stimulation). The second hypothesis is that brain components share a certain degree of similarity. This means the inner functioning of the neurons, in terms of their activation function, is similar. Excitatory and inhibitory currents differ just in their sign. Synapses of one type use the same plasticity rules throughout all layers and (modeled) areas. This should reduce the model complexity. This meets the needs that the observed complexity and differences are an emerging phenomenon of the circuit complexity and not a property of a single element. Further, the model circuit should be portable to other than the modeled brain areas. Any increase in the specificity of the mechanisms would complicate the portability.

1.8 Related models

In this section, we will introduce some related computational models and their core mechanisms. We focus on models implementing a hierarchy similar to ours and models which use Hebbian like synaptic plasticity mechanisms.

Neocognitron One of the early approaches on deep networks of the visual system was the so called “Neocognitron” (Fukushima, 1980). It consists of three stages of so called S-

layers and C-layers. A S-layer refers to the concept of feature extracting simple-cells and a C-layer refers to complex-cells which have an invariance against the stimulus position (Hubel and Wiesel, 1962). The synaptic weights of the neurons in the S-layers are learned through unsupervised learning. Each layer consists of several planes. The neurons in one plane share all the same weight matrix, but apply it on different positions. This principle is similar to a convolution, which was later used in deep convolution neural networks (e.g. LeCun et al., 1998). The weights to the S-planes are learned by using a representative neuron and its inputs to modify the weight matrix applied for all neurons in the plane. The representative neuron is selected based on its activity. This principle is also known as winner-take-all competition. The weights are increased by the input activities, which is somehow comparable to Hebbian learning. C-layers receive fixed connections from the related S-layers. A particular C-plane connects with the related S-plane only. The weights of a single neuron in a C-plane decrease monotonically with distance to the weight matrix center. Both layer types also employ inhibition within each plane. The excitatory connections to the inhibitory neurons are fixed and do decrease also monotonically with distance. The inhibitory connections to the neurons are similarly learned as the excitatory connections. The inhibition is applied divisive (shunting inhibition) in the otherwise linear activation function. The network was used to recognize different letters and digits, with different degrees of distortions. The neurons in the deepest C-layer where found to respond selectively to one stimulus pattern per plane and invariant to the stimulus position.

HMAX A model which shares a similar architecture is the "HMAX" model (Riesenhuber and Poggio, 1999). It consists of two stages of S- and C-layers, followed by a layer of so called view tuned units. Again the S-layers implement feature extraction and the C-layers produce an invariant response. However, HMAX uses no synaptic plasticity, instead the connectivity is hard wired (Riesenhuber and Poggio, 1999). The first S-layer (S1) consists of a bank of Gabor functions with different orientations and scales. The neurons in the following C-layer (C1) take the maximum response of neighboring S1 of one orientation each, but all scales. The connections of the S2-layer are arranged to represent any combination of orientations, encoded by the C1-layer. C2 makes again a spatial pooling over the S2 neurons. On the top level the view tuned units (VTU) get a task specific connectivity. Therefore, several prototypic objects (views) where presented to the network and the C2 responses where used to determine the connection matrix of the single VTUs. As acti-

vation function the neurons use a Gaussian function. Later the principle for determining the connection matrices of the VTUs was extended to the S2-layer (Serre et al., 2007). A set of stimuli was used to determine the C1 responses which again determine the connection matrix of the S2 neurons. The amount of stimuli, and consequently S2 neurons, determined the set of features encoding the objects. A further improvement to the model was to increase the sparseness of S2 inputs, giving sparser S2 responses (Mutch and Lowe, 2008). Therefore, just the most active C1 neurons of one orientation are used to determine the S2 features. Further, a mechanism inspired by lateral inhibition was used to suppress weak responses in the layers S1 and C1. Therefore, a threshold relative to the maximum response in the layer was used. All responses below this threshold are set to zero. With the HMAX model it has been shown that such an architecture is able to do invariant recognition of complex stimuli (Riesenhuber and Poggio, 1999). Later the model has been applied on typical object recognition tasks (Serre et al., 2007). Therefore, a support vector machine (linear and polynomial SVM) replaces the view tuned units. HMAX was able to outperform previous approaches for object recognition. For instance, when comparing the recognition performance, obtained with features from the model (C2 outputs), to SIFT features, which are a standard feature set for object recognition. The recognition performance can be further improved when the response characteristics and the features of the S2 units are improved with sparse features (Mutch and Lowe, 2008).

Spiking deep neural networks Similar mechanisms as in the Neocognitron or HMAX have also been implemented in spiking deep neural networks (SDNN). We focus here on networks employing unsupervised learning in terms of STDP, because of its analogy to the cortical synaptic plasticity. Masquelier and Thorpe (2007) implemented a spiking neural network consisting of two stages of S- and C-layers and an radial basis function (RBF) layer for classification on top. The input images where preprocessed by a difference of Gaussian (DoG) layer, simulating LGN on- and off-center cells. The S-layers perform a convolution on their inputs and the C-layers propagate the first spike from a spatially limited amount of afferent S-layer neurons, which mimics a max-pooling as in HMAX. The network activities are reseted to zero after each image presentation. The kernels (weights) of the first S-layer (S1) are again Gabor functions with different orientations. They are applied on different scales of the image, instead of using differently scaled kernels. The weights of the S2-layer are learned via a simple form of STDP. That is, the weights are

increased for presynaptic spikes preceding a postsynaptic spike and decreased for presynaptic spikes succeeding a postsynaptic spike. Alternatively, they tested a rate based Hebbian learning rule. Similar to the Neocognitron they used a winner-take-all strategy to select the neurons for learning, but allow two neurons to learn. For the rate based learning they used a threshold to select the learning neurons, which was defined as a fraction of the highest activity. As neuron model they used integrate-and fire neurons. In the case of the rate based learning the spike trains of the C1 neurons have been converted into a rate and a normalized response of the S2 neurons has been calculated. The RBF neurons in the top layer used a RBF function as activation function, i.e. a Gaussian as in HMAX. The RBF neurons are supervised trained to obtain a class specific selectivity. The network employs lateral inhibition in the C1 layer. The first active neuron inhibits the surrounding neurons with a distance dependent decreasing efficiency and lowers (delays) the activities. The model was evaluated on a real world classification task, where faces or motorbikes should be distinguished from background. They achieved good results with the STDP and the Hebbian learning approach and denoted the learned features informative and robust for object recognition. The authors described their approach of learning S2 features as computationally more efficient than backpropagation (the typical supervised method to learn in deep neural networks) and that it leads to less redundant features than the method used in HMAX.

A similar network design was used in Kheradpisheh et al. (2016). This network processes different image scales in parallel pathways of S1- and C1-layers which converge to a single S2-layer. The neurons in the S2-layer again learn with the previously proposed simple STDP learning rule. The responses of the C2-layer are given to a (linear) SVM as classifier, instead of using a layer of RBF units. The network was applied on a recognition task with different 3D-objects and compared to HMAX (Mutch and Lowe, 2008) and a state of the art deep convolutional neural network (DCNN) (Krizhevsky et al., 2012). The recognition task aims to measure the performance of invariant object recognition. The DCNN (DeepConvNet) was one of the best networks in the ImageNet LSVRC-2012 contest and pretrained on the ILSVRC2012 dataset. For comparison the C2 features of the model and the C2 features of HMAX have been used for classification. Further, the features of the second last fully connected layer of the DeepConvNet has been used. The model clearly outperformed HMAX and, moreover, achieved better results than DeepConvNet. Note, the DeepConvNet was trained on an entirely different dataset. However,

the features learned on the very complex ImageNet dataset are assumed to be universal enough. Conclusively, the proposed model concept was found to deal best with view point variations.

In Kheradpisheh et al. (2018) a deeper version of the previous model has been proposed. It employs an additional stage of S- and C-layers, here called Conv (convolutional) and Pool layers. As novelty all Conv-layers are learned with the previously used simple STDP learning rule. The network was again trained on images from datasets for object recognition. The learned features increased in size and complexity (Kheradpisheh et al., 2018, Fig. 2). The first layer developed edge detectors, which are roughly related to the V1 simple-cells. The second layer learns more complex composed features, like object parts. The deepest layer had features detecting object prototypes. The network was again compared to deep convolutional neural networks. That is, AlexNet, a network with good performance on the ImageNet dataset, and a network with a comparable structure to their approach. The proposed SDNN was found to have slightly better recognition accuracies than both DCNNs. However, this is just when AlexNet was not retrained on the dataset. With retraining AlexNet as supervised trained network outperforms the approach. Note, as in the previous approach the features of a deep layer from the DCNNs are used for classification. Also in comparison to a deep convolutional autoencoder with similar structure, an popular unsupervised learning method, the SDNN was superior.

VisNet Another deep rate coded architecture was proposed with "VisNet" (Wallis and Rolls, 1997). The network consists of four layers. The first layer was again modeled with predefined connectivity. A difference of oriented Gaussian was used to resemble V1 simple-cells. The oriented DoGs are defined to be selective for different orientations and frequencies, i.e. scales, and are similar to the often used Gabor functions. The neurons in the different layers are retinotop organized, each neuron was connected to a limited amount of presynaptic neurons, following a spatial organization. Not all connections are formed, a certain degree of randomness was used. A kind of local lateral inhibition was used by local contrast enhancing. Thus, all neurons learned in parallel. However, the neuron activities within a layer have been rescaled by a power law, enforcing learning of the most active neuron. As learning rule the trace learning rule of Földiák (1991) was used. For a short overview of this rule and trace learning in general please see Sec. 6.1. As natural consequence of employing trace learning temporal coherent input, containing consistent

sequences of transforming objects, was used. As stimuli few simple symbols or chars are used. In the top layer of the network some neurons developed selectivity and position invariance to these stimuli. The authors argue that the invariance is gradually achieved over the intermediate layers. Further, they tested the learning without trace and found the performance of invariant discrimination poor and comparable to a random network. When gradually modifying the trace length they found that shorter trace length than the optimal gradually decrease the performance and longer trace lengths also decrease the performance. Additionally, VisNet was trained with different faces at different positions. In contrast to the simple stimuli the top layer neurons are found to respond for more than one stimulus. The neuronal code was meant to form a distributed representation. To determine if the code conveys information on the stimuli an additional supervised trained layer (delta rule) was added to the network and the recognition accuracy was measured. With the trace rule they found a perfect classification. The performance without trace or random weights was worser, but above 80 percent. The capabilities for invariance recognition have been further tested on faces with a 3D rotation. Again a gradual increase of invariance properties in deeper layers was reported and some of the top layer neuron developed selectivity and invariance.

In Rolls and Milward (2000) a version called VisNet2 was introduced. The linear activation function and the activity normalization was replaced by a sigmoid function, for which the parameters are adjusted to get sparse responses. Further, the range of the lateral inhibition was enlarged and different modifications of the trace learning rule have been tested. These modifications have been the use of a presynaptic trace and just using the trace over the previous activations. Interestingly, it was found that using a trace not including the current activity lead to more informative neuron responses.

A further modification to the learning was introduced by Stringer et al. (2006). A Hebbian learning was successfully used to learn invariant neuron responses on simple 3D objects. The key point is that the input underlies spatially continuous transformations. That is, a large fraction of neurons encoding one view also responds for a second view. Than, a neuron responding for the first view will also respond for the second and will build up connections to the neurons just active for the second view. Interestingly, this learning scheme needs no temporal coherent input anymore, instead a kind of spatial coherence is needed. Based on the VisNet architecture multiple variations have been published. For a review please see (Rolls, 2012).

SAILnet / E-I Net A shallow spiking neural network called "SAILnet" was proposed (Zylberberg et al., 2011) to account for V1 simple-cell receptive field shapes. However, the network uses versions of rate based synaptic plasticity rules. It consists of one layer with excitatory feedforward connections and lateral inhibitory connections. The excitatory connections are learned with the Hebbian learning rule proposed by Oja (Oja, 1982). The inhibitory connections are learned with the anti-Hebbian learning rule proposed by Földiak (Földiák, 1990). The network was trained on the natural scenes dataset of Olshausen and Field (1996). With that, the neurons learned simple-cell like receptive fields. The neuron population sufficiently encoded the input images and it was shown that an input image can be reconstructed from their activities. Further, the responses of the neurons have been found to be sparse and decorrelated.

A further developed version of this network, "E-I Net", was proposed by King et al. (2013). Instead of lateral inhibition they model dedicated populations of inhibitory interneurons and excitatory neurons. The interneurons receive excitation from the excitatory neurons and inhibit them. Further, they also inhibit other inhibitory interneurons. The excitatory neurons receive natural scene input. The forward connections to the excitatory neurons are again trained with a Hebbian learning rule with Oja normalization. All connections from and to the inhibitory interneurons are trained with a novel anti-Hebbian learning rule. This rule leads to weights relative to the correlation between the neurons and is strongly related to the previously proposed rule of our group (Wiltschut and Hamker, 2009). The rule is called correlation measuring rule. The main difference to our rule is that they shifted the weight value by a baseline so that for uncorrelated neurons the weight becomes zero. With that rule they obtained simple-cell like receptive fields for the excitatory and inhibitory neurons. Again the neuron responses are found to be sparse and decorrelated. Further, they varied the amount of neurons and found that just a fraction of inhibitory neurons, around one fourth, is needed to obtain good results for image reconstruction and decorrelation. They also report that an overcomplete representation of the input gives the best results.

Own work Our group also demonstrated in shallow, rate coded, networks the capability of synaptic plasticity to learn V1 receptive fields. In Wiltschut and Hamker (2009) a one layer network with lateral inhibition and excitatory feedback to the input neurons was implemented. The input was given in terms of on- and off-center LGN responses, which

have been obtained by a whitening procedure on the input images. All connections in the network have been learned in parallel. The excitatory forward and feedback connections are learned via Hebbian learning, namely covariance learning combined with Oja normalization. The lateral inhibitory connections have been learned by anti-Hebbian learning, which aligns the weights relative to the correlation between the neurons. With that, it could be shown that simple-cell like receptive fields emerge. Further, that the strength of the inhibition influences the similarity to receptive fields from macaque monkeys. With increased similarity also measures of coding efficiency improve. For instance, population sparseness, correlation, or independence.

We advanced this work by modeling the next stage of processing in V1 by proposing a model for the development of V1 complex-cell receptive fields (Teichmann et al., 2012). Therefore, we convolved the used natural scene inputs with Gabor functions with different orientations and gave it as input to the network. The network consists of one layer of neurons with feedforward connections and lateral inhibition. The forward connections are learned with a Hebbian learning rule, which combined covariance learning, trace learning, and Oja normalization. We could show that with small temporal continuous movements in the input the most neurons developed complex-cell like invariance properties. That is, they connect to similar oriented Gabors in the input and their responses are robust to translations in the input. For further information on the learning and results please see Sec. 6.1.

Leabra Vision With Leabra Vision (LVis) a version of the Leabra framework has been proposed to model the ventral stream (O'Reilly et al., 2013). The model consists of the areas (V1, V2/V4, IT, Semantic properties, Naming output). The areas are recurrently connected with feedforward and feedback connections. Further, a inhibitory mechanism is applied within each layer. The network input was preprocessed by a DoG to resemble LGN. V1 consists of two layers. The first stage, simple-layer, applies a convolution with Gabor functions of different orientation and two frequencies. The second layer, complex-layer, applies a spatially limited max operation. V1 provide just feedforward connections to the following layer, i.e. it receives no feedback. The following areas have the full recurrent connectivity and all connections are plastic. The learning occurs by a error-driven learning rule, called XCAL[4]. While the equation of the learning appears like Hebbian learning, the protocol of the network processing and the control of the used learning threshold makes

[4] https://grey.colorado.edu/CompCogNeuro/index.php/CCNBook/Learning#Error-Driven_Learning:_Short_Time_Scale_Floating_Threshold

it similar to backpropagation (O'Reilly et al., 2013). All layers use a k-winner-takes-all competition, which selects 15 to 25 percent of the active units. This mechanism serves a uniform inhibition to all neurons, but the level of inhibition is calculated to keep the desired amount of neurons active. The Semantic properties area is trained to represent the pairwise semantic similarities between the data. The Naming output learns to represent the object class and has a more strict k-winner-take-all with k of one. The network has been trained on 3D objects underlying different transformations. The network achieved good recognition results and was demonstrated to allow robust invariant object recognition. Further, the network was tested with object occlusions. It was found that the inhibitory mechanism plays an important role in having robust responses for occluded images. A result similar to what we have shown with a single layer Hebbian network (Kermani Kolankeh et al., 2015).

Predictive Coding/Biased Competition An alternative approach to classical neural networks was proposed as "Predictive Coding/Biased Competition" (PC/BC) (Spratling, 2012). In this concept the neurons are driven by error units, which represent the global reconstruction error of the input. In a recurrent loop the neuron outputs are refined to minimize this reconstruction error. Learning in this type of model can occur by increasing weights from under represented error units to activated neurons. Spratling (2012) demonstrated a two layer network, where one layer is composed of error units recurrently connected to neurons representing the image components. He showed first on random bar input the general capability of the first layer to learn the independent components of the input. When the network was trained on natural scene images the first layer learned simple-cell like receptive fields. In the two layer network, having additional feedback connections of the second layer neuron to the first layer neuron he could show, when trained on natural scenes, that the first layer again developed simple-cell like receptive fields. The second layer learned receptive fields composed from different first layer receptive fields, i.e. combinations of Gabor-like receptive fields of different orientations (corners). This was related to receptive field properties of area V2. Later, in Spratling (2017) good performance on real world object recognition tasks have been demonstrated. However, no plasticity was used, i.e. the connections have been hard wired. Further, the feedback connection from the second layer to the first layer was removed.

2 Network Architecture

This chapter describes the connectivity and amount of our model neurons. Beyond the standard feedforward view, we implemented the connectivity based on neuroscientific data, resulting in a richer structure. The neurons in the network are retinotop organized and each neuron has a spatially limited receptive field. Neurons, which receive input from overlapping neuron distributions, are all connected to the same inhibitory neurons, which in turn inhibit these neurons. The organization of the model is scalable to any input size. But to reduce the computational costs, we have just chosen the input size so large that a V2 receptive field fits into it.

2.1 Connectivity on population level

A population comprises all neurons of the same type within the same layer in an area, e.g. all excitatory neurons in V1-layer 4. A bunch of neuroscientific data is available, describing the connectivity between these neuron populations. We derived our connectivity structure from Potjans and Diesmann (2014); Douglas and Martin (2004); Thomson and Bannister (2003). Potjans and Diesmann (2014) provided a statistical approach to the physiological data. We used the subset of highly probable connections out of their data (see Sec. 1.4). This subset should catch the main functionality of the neocortical structure. It is also consistent with the data of Douglas and Martin (2004); Thomson and Bannister (2003). For computational reasons, we are not able to implement lateral connections between the excitatory neurons. (Note, we define the term lateral connections as connections within a layer, which might differ from other usage of this term, such as all connections within a whole area.) These lateral excitatory connections lead to an instable network model, because we use additive mechanisms which allow that two excitatory neurons can drive each other and every increase of one neuron increases the drive of the other. Thus, we would obtain an exponential increase of population activities driving all neurons rapidly into saturation, losing all differentiation. In principle this issue can be overcome by reduc-

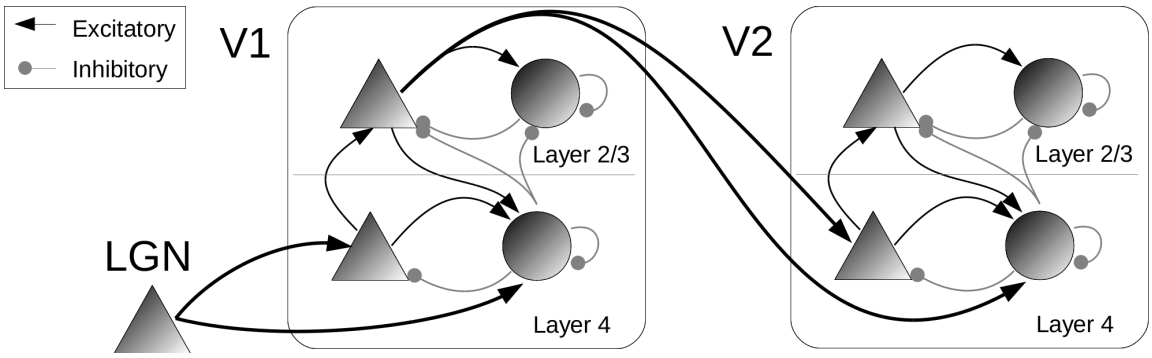

FIGURE 2.1: Model architecture. Two layer V1 model, using dedicated excitatory and inhibitory neurons. The connectivity base on neuroscientific foundations.

ing the influence of the lateral excitatory connections, so that they can not drive neurons solely under typical stimulation conditions. Indeed, than this connections lose their relevance. Alternatively, the inhibitory circuit has to be able to balance the excitation. The limitation of the linear mechanism, employed by us, could be potentially overcome by using a nonlinear (super-linear) activation function in the inhibitory neurons. We also do not regard the layers 5 and 6, as they serve the cortico-thalamic-cortical connection structure, which we do not use in this thesis. The used connection data cover mainly V1. For V2, we used the naive approach that this area shares a similar structure as V1. This assumption is covered by the data of Douglas and Martin (2004), which described the main connectivity between the populations in the neocortex as similar. Figure 2.1 provides an illustration of the connections between all neuron populations of our network.

Our model consists of the areas V1 and V2, receiving input from area LGN. LGN, consists just of excitatory neurons, solely driven by the input. For V1 and V2, we implemented the most important layers, layer-4 and layer-2/3. Each cortical layer again consists of two neuron populations, resembling the two main categories of cortical neurons, inhibitory and excitatory neurons. We set the input of the model as activity on neurons of a so called IN-layer. The neurons of this layer have no dynamic and just represent the image values. Each IN neuron projects to one LGN neuron. LGN is the first area with neural dynamics, the neurons follow the same activation function as all further neurons of the model (see Sec. 5.2). This means, their firing rate is defined as differential equation following the stimulating currents. We did not adapt the the intrinsic parameters of LGN. The first plastic connections emerge from LGN to the excitatory and inhibitory populations of V1-layer 4 (V1-L4). Like all inter area connections, the inter area connections in our model emanate

Layer	Type	Geometry	Neurons
IN	Input	24x24x2	1152
LGN	Excitatory	24x24x2	1152
V1-L4	Excitatory	13x13x4	676
	Inhibitory	13x13x1	169
V1-L2/3	Excitatory	7x7x12	588
	Inhibitory	7x7x3	147
V2-L4	Excitatory	4x4x20	320
	Inhibitory	4x4x5	80
V2-L2/3	Excitatory	2x2x75	300
	Inhibitory	1x1x75	75

TABLE 2.1: Layer geometry and amount of neurons. Overview on the organization and amount of neurons of a certain type in the different layers.

from the excitatory neurons of the preceding area and target all neurons in the receiving layer, i.e. excitatory and inhibitory neurons (Isaacson and Scanziani, 2011). Similarly to LGN, the excitatory neurons of V1-layer 2/3 (V1-L2/3) are connected to all neurons of V2-layer 4 (V2-L4). Within each cortical layer, the excitatory population serves input to the inhibitory population and the inhibitory neurons are inhibiting the excitatory neurons. This architecture resembles the interplay between excitation and inhibition within the visual cortex. The inhibitory neurons learn to inhibit similarly responding excitatory neurons and enforce, in the learning phase, the differentiation of these neurons. Another benefit of such a lateral inhibition mechanism is that inhibitory connections can store typical competitors of the represented patterns and the suppression of neurons responding for similar patterns can enhance the robustness to distortions of the input (Kermani Kolankeh et al., 2015). In our model, multiple inhibitory neurons can have a similar initial connectivity. If these neurons would solely receive inputs from the excitatory neurons, than all neurons would tend to learn similar patterns. To facilitate differentiation within a population of inhibitory neurons, neighboring inhibitory neurons inhibit each other. Further, the neurons of the layers 4 project to layer 2/3 neurons and provide thereby the classical feedforward stream in the hierarchical processing. For a full overview about all connections within our model please see Figure 2.1 and also Table 2.4.

2.2 Population geometry

The geometry is the amount and arrangement of the neurons. The population geometry is designed based on the principles of retinotop organization, the ratio between excitatory and inhibitory neurons, and also considers the ratio between the amount of neurons in the different layers. Therefore, we used the data of Potjans and Diesmann (2014) about the amount of neurons within one square millimeter of the primary visual cortex (V1). As input size, we used 24 by 24 by 2, where the third dimension stands for the cell type of on-center and off-center cells. The size of 24 by 24 is chosen more than twice as large as a typical receptive field of a V1-L4 neuron. This should ensure that V2 neurons can have receptive field sizes twice as large as V1 neurons, which is related to the increase of the receptive field sizes in the visual cortex. We defined that the initial receptive field size of a V1-L4 neuron should cover the half input size (12 by 12). In previous studies, we obtained good results with that size (Wiltschut and Hamker, 2009). The initial receptive fields of neighboring V1-L4 neurons lay next to each other with a shift of one. The neighborhood is defined in the first two dimensions of the layer geometry. As a consequence of the 24 by 24 LGN size and the size of the initial receptive field of the neurons, the sizes of the subsequent layers have to be smaller, so that the neurons at the borders can have the same receptive field size as all other neurons. This means, that the x-y size of the V1-L4 layer has to be chosen 13 by 13. The third dimension (z) scales the amount of units being at the same position in the x-y grid. For V1-L4, we chose four as amount (z) of the excitatory neurons and one for the inhibitory neurons, to save computational costs. The subsequent layer sizes are chosen so that each neuron receives input from a bit more than the half of the previous layer. In the primate visual cortex the amount of neurons in area V2 is smaller than in V1. We approximately halve the amount of neurons in V2, which is a stronger reduction as observed in the visual cortex (DiCarlo et al., 2012). An overview on the amount of neurons in each network layer can be found in Table 2.1 and the initial receptive field sizes in Table 2.4.

As mentioned before, we have chosen the ratios between the amount of neurons within the populations based on the data of Potjans and Diesmann (2014). The authors also reported the amount per square millimeter V1 surface (Table 2.2). The ratio between inhibitory and excitatory neurons is approximately one fourth and layer-2/3 has slightly fewer neurons than layer-4. We picked the third dimension to get a ratio between inhibitory and excitatory neurons of one to four. The comparison of the ratios to the data of V1 are

Layer	Type	Amount	Ratio to L4-Exc	Ratio to L2/3-Exc
L4	Excitatory	21915	**1**	1.0596
	Inhibitory	5479	**0.25**	0.2649
L2/3	Excitatory	20683	**0.9438**	1
	Inhibitory	5834	0.2662	**0.2821**

TABLE 2.2: Ratio between the amount of neurons of different populations in V1. The amount is given for one square millimeter cortex. The ratios are reported in relation to the amount of excitatory neurons of layer-4 (fourth column), respectively layer-2/3 (last column). Bold numbers indicate the important values for the ratio between excitatory and inhibitory neurons within a layer and the ratio between the excitatory neurons of layer-4 in relation to layer-2/3. (Data taken from Potjans and Diesmann, 2014)

sufficiently close (Tab. 2.3). V1-L4 fits the data perfectly, V2-L4 as well. We aligned the layers 2/3 to the related layer-4 and found that V1-L2/3 has 8 percent fewer excitatory neurons than the ratio from the V1 data would give. The excitatory neurons of V2-L2/3 matched the data well. However, the inhibitory neurons of V1-L2/3 and V2-L2/3 are 11 percent fewer compared to the data.

Layer	Type	Amount	Ratio to L4-Exc	Ratio to L2/3-Exc
V1-L4	Excitatory	676	**1**	1.0612
	Inhibitory	169	**0.25**	0.2653
V1-L2/3	Excitatory	588	**0.8698**	1
	Inhibitory	147	0.2175	**0.25**
V2-L4	Excitatory	320	**1**	1.0667
	Inhibitory	80	**0.25**	0.2667
V2-L2/3	Excitatory	300	**0.9375**	1
	Inhibitory	75	0.2344	**0.25**

TABLE 2.3: Ratio between the amount of neurons of different network populations. The ratios are reported in relation to the amount of excitatory neurons of layer-4 (fourth column) or layer-2/3 (last column), of an area. Bold numbers indicate the important values for the ratio between excitatory and inhibitory neurons within a layer and the ratio between the excitatory neurons of layer-4 in relation to layer-2/3.

Target Layer	Neuron Type	Source Layer	Neuron Type	Init RF Size	Offset
IN	Input	-	-	-	-
LGN	Excitatory	IN	Excitatory	1	-
V1-L4	Excitatory	LGN	Excitatory	12x12x2	0
		V1-L4	Inhibitory	11x11x1	-5
	Inhibitory	LGN	Excitatory	12x12x2	0
		V1-L4	Excitatory	11x11x4	-5
		V1-L2/3	Excitatory	7x7x12	-6
		V1-L4	Inhibitory	11x11x1	-5
V1-L2/3	Excitatory	V1-L4	Excitatory	7x7x4	0
		V1-L4	Inhibitory	11x11x1	-5
		V1-L2/3	Inhibitory	5x5x3	-2
	Inhibitory	V1-L2/3	Excitatory	5x5x12	-2
		V1-L4	Inhibitory	11x11x1	-5
		V1-L2/3	Inhibitory	5x5x3	-2
V2-L4	Excitatory	V1-L2/3	Excitatory	4x4x12	0
		V2-L4	Inhibitory	3x3x5	-1
	Inhibitory	V1-L2/3	Excitatory	4x4x12	0
		V2-L4	Excitatory	3x3x20	-1
		V2-L2/3	Excitatory	fully con.	-
		V2-L4	Inhibitory	3x3x5	-1
V2-L2/3	Excitatory	V2-L4	Excitatory	3x3x20	0
		V2-L4	Inhibitory	3x3x5	-1
		V2-L2/3	Inhibitory	fully con.	-
	Inhibitory	V2-L2/3	Excitatory	2x2x75	-1
		V2-L4	Inhibitory	3x3x5	-1
		V2-L2/3	Inhibitory	fully con.	-

TABLE 2.4: Connectivity and initial receptive field sizes. Overview on the initial connection windows (receptive field sizes) for each neuron type in each layer. The connection type is identical with the neuron type of the source layer.

2.3 Receptive field sizes and retinotop organization

The neurons initial receptive field sizes are determined based their retinotop organization. Each single neuron is only connected to a patch of the neurons from the afferent population. This patch is determined based on the position of a neuron in the x-y grid of its population and has a rectangular shape. For instance, the first excitatory neurons at the positions 1-1-1 to 1-1-4 of V1-L4 are all connected to the first neurons of their afferent population LGN, which means to the neurons 1-1-1 to 12-12-2. This is because their receptive field size is 12 by 12. Within a population, neurons have been connected to their neighboring neurons. We shifted the rectangular patch by an offset in x and y direction to center over the neuron itself. This resulted in a symmetric connectivity to all neighboring neurons.

Following pseudo code is describing the algorithm (Algorithm 2.1). The code uses the typical Python and ANNarchy 3 commands. Arguments are the pre and post populations and the connection type, to identify the populations, which should be connected, and to give the connection an type identifier. The connection type can be excitatory or inhibitory and is linked to the respective learning rule. Further parameters are, the min and max value, between the initial connection strength is drawn. Followed by the arguments for the receptive field width and height of a single neuron, e.g. 12 by 12 for the LGN to V1-L4 connections, the distance to the receptive field of next postsynaptic neuron, the x and y offset for the displacement, and the neurons connection delay in Euler steps.

First, the algorithm allocates data structures, containing the identifiers of the presynaptic neurons (*rank_list*), the weight values which store the connection strength (*value_list*), and the connection delay (*delay_list*). The delay can in principle be chosen individual for each connection, which is not used by this connection pattern. We used the same delay of one Euler step (1ms) for all connections, so that each element of the *delay_list* is allocated with the value parameter *delay* and will remain unchanged in the algorithm. Second, we determine for the receptive field of each postsynaptic neurons the start and end neuron in the presynaptic layer in x and y dimension. The first neuron in the presynaptic layer is calculated as the position *x_post* (*y_post*) in the postsynaptic layer, multiplied with the shift rf_shift between the receptive fields, plus the offset $offset\_x$ ($offset\_y$), and is clipped at the boundaries of the presynaptic population. The calculation of the upper boundary is calculated analogue, just the rf_width (rf_height) is added. We determine just the x and y boundaries and create connections to all presynaptic neurons in the depth plane. Note, that we are precluding self connections. Third, for all presynaptic neurons, where we

want to form a connection, the neuron identifier, called rank, is determined and added to the *rank_list*. The weight value is drawn from a uniform distribution, between *min_value* and *max_value* and added to the *value_list*. When the values for all connections of one postsynaptic neuron are determined, the connect procedure of ANNarchy is called and the connections are created. Depending on the layer geometry and the chosen offset, the neurons at the borders of a population can get different receptive field sizes.

The parameters which we have used for connections between the populations are given in Table 2.4. As mentioned before, the excitatory in V1-L4, as well as the inhibitory neurons, receive input from a 12 by 12 patch of LGN and see the half (in one extension) of the total inputs initially. The excitatory neurons in V1-L2/3, receive from approximately the half of the V1-L4 neurons input and have access to up to two thirds ($7+12-1=18$) of the network input. The excitatory neurons in V2-L4 again, see approximately the half of the V1-L2/3 neurons and see seven eighths ($4+18-1=21$) of the input. A single excitatory neuron in the highest layer V2-L2/3, sees three quarters of the excitatory neurons of V2-L4 and, thus, nearly the full input ($3+21-1=23$). Negative offsets are used to shift the receptive fields so that all neurons of the presynaptic population are symmetrically covered. This is important for the connections of the inhibitory populations to themselves. Without an offset the last neuron of a population would have just connections to itself, which we do not allow. This also allows to define connections to all neighboring neurons in a certain range. We set the rf_shift parameter for all connections to one. All here described receptive field sizes are used for initially connecting the network neurons. Within the learning phase structural plasticity (Chapter 4) is continuously changing the connection structure, based on the learned synaptic weights. Thus, the initial connection structure do not limit the variety of receptive fields shapes and sizes.

```
def connect_populations(population_pre, population_post, connection_type, min_value,
  max_value, rf_width, rf_height, rf_shift, offset_x, offset_y, delay)
  #Preallocate (maximal) lists
  rank_list  = (rf_width * rf_height * population_pre.depth) * [0]
  value_list = (rf_width * rf_height * population_pre.depth) * [0]
  delay_list = (rf_width * rf_height * population_pre.depth) * [delay]

  #Connect all postsynaptic neurons
  for x_post in xrange(population_post.width):
    for y_post in xrange(population_post.height):
      #Determine x-y coordinates of presynaptic neurons based on post neuron position
      lowerbound_x = max(x_post * rf_shift + offset_x, 0)
      upperbound_x = min(x_post * rf_shift + offset_x + rf_width, population_pre.width)
      lowerbound_y = max(y_post * rf_shift + offset_y, 0)
      upperbound_y = min(y_post * rf_shift + offset_y + rf_height, population_pre.height)

      for z_post in xrange(population_post.depth):
        #Determine ranks and weight values for all connected presynaptic neurons
        nb_connections = 0
        for x_pre in xrange(lowerbound_x,upperbound_x):
          for y_pre in xrange(lowerbound_y,upperbound_y):
            for z_pre in xrange(population_pre.depth):
              #No self connections
              if not ((population_pre == population_post) and (x_post == x_pre) and
                (y_post == y_pre) and (z_post == z_pre)):
                #Add neuron rank to rank list
                rank_list[nb_connections] =
                  population_pre.rank_from_coordinates(x_pre,y_pre,z_pre)
                #Add random weight value to value list
                value_list[nb_connections] = random.uniform(min_value, max_value)
                nb_connections = nb_connections + 1

        #Add connections to the neuron
        post_neuron = population_post.neuron(x_post,y_post,z_post)
        post_neuron.connect(population_pre, rank_list[0:nb_connections],
          connection_type, value_list[0:nb_connections], delay_list[0:nb_connections] )
```

ALGORITHM 2.1: Pseudo code of the connection algorithm. After preallocating the data structures, the connections to the postsynaptic neurons are determined. Each postsynaptic neuron is connected to a patch of presynaptic neurons. This patch is determined based on the x-y position of the postsynaptic neuron and its offset. If the list of the ranks of the presynaptic neurons and the weight values are completed, the connections are added to the postsynaptic neuron.

3 Network Training and Evaluation

This chapter introduces the input stimuli and the presentation protocol for network training. We describe the preprocessing of the input data and how test data are processed. Further, we introduce the used software for simulating and evaluating the network model.

3.1 Neural simulator and evaluation software

To simulate the network we used the neural simulator *ANNarchy* (Vitay et al., 2015) in version 3.09 with extensions for structural plasticity. The newest simulator version is online available[1]. In the used version the neuron model and synaptic plasticity are defined in C++ language. The description of the network structure is defined via a Python 2.7 interface to the neural simulator. Structural plasticity is implemented in Python over the simulator interface. Also the network input is provided over this Python interface as well as the control of the presentation procedure and the access and recording of the network activity. For the network evaluation we recorded the network states and response data within the neural simulator. For further analyzes we load these data into *MATLAB*[2] and calculated measures and graphs.

3.2 Training data

Goal of the network is to learn a model of the early visual system with a realistic receptive field structure. It was found that with natural scenes as input V1 simple-cell receptive fields emerge (Olshausen and Field, 1996). Thus, we used the natural scene dataset of Bruno Olshausen[3] for network training. The dataset consists of 10 monochrome images with a resolution of 512 by 512 pixels. This dataset has been successfully used in many

[1] `https://bitbucket.org/annarchy/annarchy`
[2] `https://de.mathworks.com/products/matlab.html`
[3] `http://redwood.berkeley.edu/bruno/sparsenet/`

different models for simple-cell receptive field learning (e.g. Olshausen and Field, 1996, 1997; Hoyer, 2004; Rehn and Sommer, 2007; Wiltschut and Hamker, 2009; Clopath et al., 2010; Zylberberg et al., 2011; Spratling, 2012; Teichmann et al., 2012). Note, when we refer to *natural scenes* used in our model, we mean preprocessed patches from this dataset.

3.3 Preprocessing

Because of large inequities in the variance of the images and the power spectrum Olshausen and Field (1996) pre-whitened their image data. They applied a circular symmetric low-pass frequency filter in the Fourier domain which should equalize the amplitudes in the spectrum, but avoid an amplification of noise and artifacts from the rectangular images. They chose following exponential function with a cutoff frequency $f_0 = 200$ and a steepness of $n = 4$ (Eqn. 3.1). For a more detailed description see Olshausen and Field (1997).

$$R(f) = fe^{-(f/f_0)^n} \tag{3.1}$$

When applying this filter on the images the resulting images have an approximately symmetric distribution of positive and negative values. When projecting the used filter into image space the shape of the filter is similar to the center surround structure of LGN receptive fields (Fig. 6 in Olshausen and Field, 1997). Thus, when applied on an image than the positive valued results can be interpreted as responses of an on-center LGN neuron and negative values as the response of an off-center neuron. Accordingly, we mapped the positive responses to the first entry in the z-dimension of a three dimensional matrix representing our image and the absolute values of the negative responses to the second entry of the z-dimension. At all positions with no value of the respective category we filled the matrix with zeros, giving us a code with zero entries for the off-center LGN responses at positions where an on-center LGN cell responds and vice versa.

Finally, we normalize the values in the matrix to obtain a suitable neuron activity in the range [0,1]. Therefore, we divide the resulting matrix by its 99.9th percentile value. This gives us approximately the desired range as just 0.1 percent of the data have higher values, which is that rare that we have no pixel has values higher than one in the most image patches.

We applied the filter, mapping, and normalization steps on all input images we present as input to our network. We do this image wise so that the range of each image is similar,

but can differ between patches.

3.4 Presentation protocol

For the training of our network we have to select patches with the size of our input population from the preprocessed natural scenes. The values of these patches are set as input on the corresponding neurons of our first network layer, named LGN. The model LGN neurons have the same activation function as all other model neurons (see Sec. 5.2), but without any adaptation of the parameters (see App. A.2). Thus, they implement a differential equation for the firing rate following the input strength.

3.4.1 Selection of image patches

Our model should be enabled to learn invariance properties via trace learning. Therefore, we need a temporal continuous stream of input images with small changes in the position of the image content. Similarly to Teichmann et al. (2012) we simulated micro-movements inspired by fixational eye movements (Martinez-Conde et al., 2004; Rolfs, 2009). These movements are implemented as random walk across the image. The algorithm runs as following:

1. we select randomly an image
2. we determine randomly the position of an image patch, this patch is present to the network for a certain amount of time
3. we calculate a new slightly changed position of the patch and present it to the network
4. we slightly change the position ten times before starting with the first step again.

Note, the size of a patch is much smaller than the image size. In the rare case that we reach the border of the image we select another new position of the patch. In Teichmann et al. (2012) we controlled the stability of the trace learning, with the same trace length, for different amounts of consecutive images. We found that the trace learning was also functioning with just five consecutive images before a new random image and patch is drawn. Because of the random selection of the image patch we obtain an equal statistic

Distance	Probability in percent
1 pixel	44.04
2 pixel	30.27
3 pixel	16.20
4 pixel	6.75
5 pixel	2.19
6 pixel	0.55

TABLE 3.1: Distances and their probability for the input patch shift. The probabilities follow a Gaussian with $\sigma = 2$ and $\mu = 0$.

for each image pixel in the input, which is important for the stability of the learning. We determine the subsequent patch position with the following algorithm, which is a slightly modified version from Teichmann et al. (2012). First, we randomly select an angle out of 360 degree for the direction of the movement. Than we randomly select a distance. The distance is drawn from one to six pixels with decreasing probability following a Gaussian distribution with $\sigma = 2$ and $\mu = 0$ (see Table 3.1). Finally, we select coordinates of the patch at the next discrete position to the target of the micro-movement.

3.4.2 Presentation time and stimuli amount

We present each stimulus for 100ms (*presentation time*). The differential equations of the network are evaluated every millisecond. This is sufficient long that the activity of the network has converged. We present 500000 natural scene stimuli to the network, which gives us 50 million network steps to calculate. The network converges before one fifth of the training time of 100000 stimulus presentations (see Sec. 4.3.5).

3.5 Network initialization

In the initial state of the network all neurons have zero activity and membrane potential. All weights are positive valued and drawn from a uniform random distribution. The expectation values of the excitatory weights are chosen in that way that all excitatory connections together a neuron receives induce excitatory current with about 0.5 as expectation value. The inhibitory connections are initialized with a smaller random range, sufficient strong to sparsify the population responses. Because of that the activities of the neurons should be sufficiently differentiated which facilitates a fast differentiation of the neurons receptive

fields.

3.6 Stimulation protocol for evaluations

To any input stimuli we present to our network we apply the described preprocessing to obtain the whitened on-off-center structure for our input. We always apply the preprocessing with the same parameters, which we found suitable for all used image datasets. After the network training we turn off all plasticities (structural, intrinsic, and synaptic plasticity) before we present test stimuli to the network. Further, we reset all network activities and membrane potentials to zero before we present a stimulus. We record the responses 100ms after the stimulus onset (10 times the time constant of the neurons activation function). Any further details about the stimuli and presentation protocols are described in the respective evaluation sections.

4 Structural Plasticity

This chapter introduces the structural plasticity mechanisms. First, we describe the related computational models and their mechanisms. Then, we present our implementation. Subsequently, we evaluate its functioning and stability. An initial version of the presented implementation has been developed together with the exchange student Maxwell Shinn.

4.1 Introduction

One challenge in creating neural models of the visual system is the appropriate definition of the connectivity. The modeler constrains the results with its definition, also for learning models. Using too few connections neurons will not develop appropriate receptive fields. Using too many the model might lose features like retinotopic organization. Further, often the precise knowledge about appropriate connection sizes is lacked, for instance in deeper layers of the cortex or for different neuron types like interneurons. Also within the same population of neurons receptive field sizes can largely differ. Hence, a mechanism at hand refining the connection structure based on the learned weights would be appreciated. Such a mechanism can be found in the human brain by structural plasticity.

In recent years several models of structural plasticity have been developed. These models largely differ in the used methods. However, the most models treat structural changes as random process. Nevertheless, several models link the process of synapse removal directly or indirectly to the strength of the synaptic weight. Zheng et al. (2013) developed a self-organizing recurrent network with a simple form of STDP. A synapse is removed (non random) when the weight becomes, through synaptic plasticity, lower than zero. Similarly, in Deger et al. (2018) a synapse is removed when, through STDP, the weight reaches zero, however, two neurons can have multiple contacts so that a connection is finally removed when the total weight becomes zero. Fauth et al. (2015b) implemented a two neuron model to investigate the relation between synaptic plasticity and structural plasticity. There synapse removal has been defined as probabilistic process depending on the strength of the

synaptic weight. That is, synapses are removed using a sigmoidal probability function with high values for weak weights.

Beside a direct weight strength dependent removal some models indirectly rely on the weight strength. Helias (2008) implemented a network of integrate-and-fire neurons which learns the excitatory connectivity with a plasticity rule accounting for LTP. Structural plasticity acts here as counterpart to LTP and removes synapses if the correlation between the pre- and postsynaptic neuron's activity drops below a threshold. Alternatively, in Deger et al. (2012) a sigmoidal shaped probability function is used for the removal, if the correlation is below average. Similarly to Hebbian learning, correlation bases on the coactivity of the pre- and postsynaptic neuron. Thus, basing the synapse removal on the correlation between the neuron's activity is strongly related to a weight strength based approach. Further, strong excitatory synaptic connections would necessarily lead to highly correlated neuronal activity.

The process of synapse removal has also been explained by a model using random fluctuations in the spine volume (Yasumatsu et al., 2008). It has been shown that the finding can be reproduced that thin spines are much more likely to disappear than thick spines. Another criterion, independent from individual synaptic weights, is the homeostasis of the neuron's activity. It was used to change the amount of dendritic or axonal contacts a neuron can form (Butz and van Ooyen, 2013; Butz et al., 2014a,b; Gallinaro and Rotter, 2018). If the amount of these contacts drops below the actual number of formed synapses a random synapse is removed. Much simpler, the constant decay of the synapse amount through random removal has been successfully used (Butz et al., 2009a; Fauth et al., 2015a).

As well as for synapse removal constant probabilities are often used for the decision whether a synapse is formed (Deger et al., 2012; Zheng et al., 2013; Deger et al., 2018). Fauth et al. (2015a,b) used a constant probability to form synapses out of a set of potential synapses between two neurons. However, relying on the definition of potential synapses the model brings, implicitly, a spatial restriction where new synapses can be formed. That is, new synapses can just be formed where axons are close to the dendrite of a neuron. Regarding the location of neurons in the cortical tissue and assuming axons and dendrites are growing outgoing from this location the vicinity between these neurons can be taken into account. Butz et al. (2009a) used a Gaussian distance measure within a grid of neurons to determine the formation probability. This approach has been extended by considering the amount of vacant synaptic elements (Butz and van Ooyen, 2013; Butz et al., 2014a,b).

The amount of these elements is changed based on activity homeostasis. That is, increasing the amount when the neuron is active below a target value and decreasing it if it is active above the target value. Note, that here the synaptic weights are constant. Similarly, but without considering the distance between the neurons, Gallinaro and Rotter (2018) formed randomly after fixed time intervals new synapses when neurons have vacant elements. Surprisingly, the model of Helias (2008) comes with no synapse formation. Each neuron starts there with k excitatory synapses which are removed when the pre- and postsynaptic neuron's correlation is low. Thus, just connections between highly correlated neurons remain. Indeed, this method seems to be ineffective for any noisy system where for longer time periods correlations can be low. Then this approach would lead to a nearly total wipe out of synapses, where just few remaining connections dominate the postsynaptic activity. It also do not match findings where the majority of found cortical synapses have just a low contribution to the postsynaptic activity (e.g. Cossell et al., 2015).

We assume for our model that structural plasticity is a stochastic process without any guidance. Synapses are created in the vicinity of existing synapses and weak synapses are removed. Moreover, we aim to reduce the influence of the initial connectivity on the learnings of the network. The outcome of the model should be experience driven. Initially, we use retinotop organized connections, which is a desired property of the visual cortex. The structural plasticity should not impair this organization, which would be likely for a stochastic process. We assume that the synapse creation in close range to existing synapses prevents the system from implausible expansions of the receptive fields to spatially distant locations. The method should be usable for large connection matrices, where a fully stochastic process would lead to small probabilities for forming synapses close to existing ones. Beyond that, it should be applicable on all types of synapses. We assume that a stochastic process just utilizing the vicinity of the synapses as additional criterion to the weight strength will work for excitatory and inhibitory synapses. We describe the mechanisms and assumptions of our implementation in Sec. 4.2. We apply structural plasticity on all model synapses, similar as we apply synaptic plasticity on all synapses in parallel. Subsequently, we evaluate the changes of the connectivity induced by the mechanisms (Sec. 4.3). That is, the amount synapses and the receptive field sizes. We compare the development of the receptive fields from different initial conditions. We also control if the retinotop organization of the receptive fields is preserved and the network convergence is not impaired.

4.2 Experience-dependent spatial growth model

In our model we have no concrete representation of the dendritic tree or the axonal branches. However, our neurons are initially retinotop organized and the connections between two layers are described by a set of pairs of presynaptic and postsynaptic neurons. Regarding a single postsynaptic neuron, this gives a weight matrix with the retinotop layout of the presynaptic layer. Thus, we can locate our synapses in the coordinate system of the afferent layer. Despite this, we lack a concrete representation of the position of the axonal-dentrite contact. However, we can make the likely assumption that connections to neighboring presynaptic neurons would be in immediate vicinity of the regarded synapse. Hence, we define the neighboring synapses in the retinotop grid as potential synapses, being likely to be formed in the case of a spine outgrowth. Further, it has been found that it is likely that new synapses are formed in the vicinity of existing strong synapses (Caroni et al., 2012; Harvey et al., 2008; De Roo et al., 2008). Thus, we exploit the spatial synapse organization of our model to determine a probability for creating a synaptic connection based on the strength of neighboring synapses.

New formed synapses are originated by thin spines (Knott et al., 2006). Such spines are likely to fast disappear (Yasumatsu et al., 2008), spines with larger volume are found more stable (Knott et al., 2006; Yasumatsu et al., 2008). New spines are subjected to normal synaptic plasticity promoting stabilization (Caroni et al., 2012; Holtmaat et al., 2005), which acts much faster than structural changes. Therefore, newly created synapses in our model begin with a small randomly chosen synaptic weight value (see 4.2.1). The value is determined with an expectation value equal to the value where synapses are eliminated with the half of the maximum delete probability. This process is based on foundations that new formed synapses having mostly similar synaptic volumes as synapses which are likely to disappear (e.g. Yasumatsu et al., 2008).

Whereas the formation of new synapses seems to depend on the synaptic strengths in the neighboring (Caroni et al., 2012; Harvey et al., 2008; De Roo et al., 2008), synapse elimination is closely related to the volume of a synaptic spine (Yasumatsu et al., 2008). Thin spines are more likely to disappear than thick ones, which can be stable for long periods (Knott et al., 2006; Kasai et al., 2010; Yasumatsu et al., 2008). We relate the deletion probability of synapses to the weight value (see 4.2.2) (Fauth et al., 2015b; Yasumatsu et al., 2008; Butz et al., 2009b). Low weights are likely to be removed whereas high one are subjected to be very stable (Yasumatsu et al., 2008).

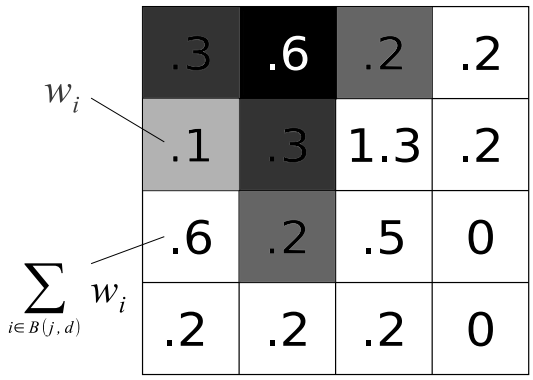

FIGURE 4.1: Build probabilities around existing synapses. Here the sum of the neighboring weights, mainly determining the build probability, is illustrated. Values in gray shaded boxes denote the weight strength of existing synapses. White boxes denote non-existing synapses, having the sum of neighboring weights inside. The illustration uses a neighborhood distance d of one. Non-existing synapses surrounded by strong synapses will have higher build probabilities than others.

4.2.1 Synapse creation

First we define the so called build probability p_j^{build} for a non-existing synapse j (Eqn. 4.1). This probability describes how likely it is that a synapse is created within a particular time interval (see Sec. 4.2.3). This is determined based on the sum of the synaptic weights w_i of the existing synapses i in the neighborhood $B(j,d)$ (Fig. 4.1), normalized by the maximum weight w_k of the postsynaptic neuron's weight matrix and divided by the size of its neighborhood $|B(j,d)|$. The result is scaled by the constant c_s. The neighborhood $B(j,d)$ is defined as set of all synapses around synapse j, within a range d in any dimension of the population grid.

$$p_j^{build} = c_s \cdot \frac{1}{|B(j,d)|} \cdot \sum_{i \in B(j,d)} \frac{w_i}{\max_k(w_k)} \tag{4.1}$$

Regarding a very large population of afferent neurons and a sparse connection structure, the calculation of the build probability for each non-existing synapse (connection) is computationally inefficient. Thus, the equation is transformed to calculate just values in the neighborhood of existing synapses (Eqn. 4.2). Therefore, we calculate the probability incrementally by accumulating the normalized weights of the existing synapses i onto the non-existing synapses in their neighborhood. So we can write the probability p_j^{build} as sum of all synaptic weights in the neighborhood $B(j,d)$, where each $(\forall i \in B(j,d))$ synaptic strength w_i is weighted by its neighborhood size $|B(i,d)|$. This weighting leads to a negli-

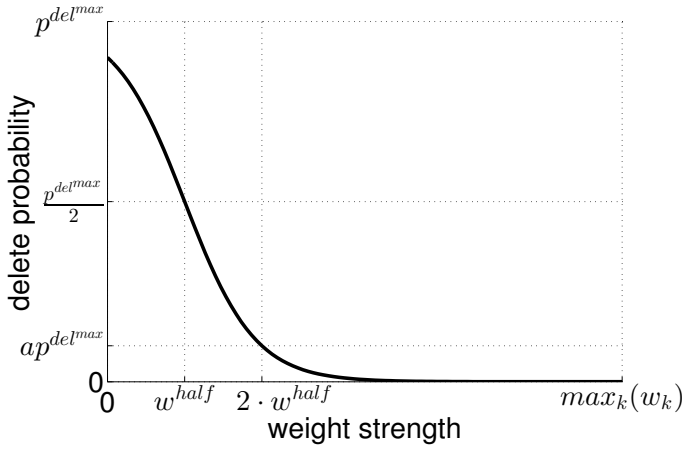

FIGURE 4.2: Delete probability function with parameters. The function drops from its extreme value, the maximal delete probability $p^{del^{max}}$, close to zero. For weights with weight strength w^{half} the deletion probability is half of its maximum and for twice this weight strength its a times the maximal probability.

gible difference at the borders of a population, where the neighborhood sizes can differ.

$$p_j^{build} = c_s \cdot \sum_{i \in B(j,d)} \frac{1}{|B(i,d)|} \frac{w_i}{\max_k(w_k)} \tag{4.2}$$

With this definition at hand, we can calculate the p^{build} values for all relevant non-existing synapses by iterating over the existing synapses i and increasing the build probabilities in their neighborhood (Eqn. 4.3).

$$p^{build}(B(i,d)) = p^{build}(B(i,d)) + c_s \cdot \frac{1}{|B(i,d)|} \frac{w_i}{\max_k(w_k)} \qquad \forall i \tag{4.3}$$

If a new synapse is created, its weight w_j is chosen from a uniform distribution between zero and twice the value w^{half} multiplied with the maximum weight $max_k(w_k)$ of the neuron. This places the weight value around the value where the deletion probability is half its maximum. The used parameters are listed in Table A.1.

4.2.2 Synapse removal

As well as for the formation of synapses we treat the removal as probabilistic process. Contrarily, we base the probability p_i^{del} for deleting an existing synapse i only on the weight strength w_i of the synapse itself (Eqn. 4.4). The probability follows a logistic function, having its maximum for low weight values to remove weak synapses with higher probability (Fig. 4.2). If the normalized weight strength increases to w^{half} the removal probability is decreased to the half of its maximal value $p^{del^{max}}$. For two times w^{half} the removal

probability drops to the fraction a of its maximum. Hence, with the parameters w^{half} and a the shape of the removal probability function can be adjusted. Where w^{half} defines the setpoint between high and low removal probabilities and a controls the steepness of the transition, i.e. for very low values we get a sharp transient and a smooth one for higher values.

$$p_i^{del} = \frac{p^{del^{max}}}{1 + e^{\frac{\frac{w_i}{\max_k(w_k)} - w^{half}}{\Delta}}} \tag{4.4}$$

The parameter Δ (Eqn. 4.5) is defined so that the parameter a in relation to w^{half} sets the result of equation 4.4 for $w_i = 2 \cdot w^{half}$ to $a \cdot p^{del^{max}}$.

$$\Delta = \frac{w^{half}}{\ln\left(\frac{1}{a} - 1\right)} \tag{4.5}$$

The used parameters are listed in Table A.2.

4.2.3 Probability calculation

The probabilities for synapse creation or removal $p^{build|del}$ are defined for a fixed fine grained interval of $t^{base} = 1s$ to allow a very smooth transition of network structures. Indeed, structural plasticity is a very slow process in comparison to synaptic or intrinsic plasticity. Hence, to save computational costs, we update the connection structures just every $\Delta t = 20s$. Note, one network step has $1ms$. Thus, we have to calculate how likely it is that a synapse is created or removed with the given probability after an interval of Δt.

The probability for the opposing event that a synapse is not changed (formed or removed) is given by one minus the build or removal probability. We define that a synapse is changing after Δt if the synapse is changed in one of the base intervals since the last update. We ignore the possibility that deleted synapses can be recreated and vice versa, which is suitable for low formation and removal probabilities. Hence, we get that a synapse will not change if the synapse has not changed within each of the previous base intervals. Thus, a synapse is changing with the probability of the opposing event (Eqn. 4.6).

$$
\begin{aligned}
P(\text{synapse not changed}) &= 1 - p^{build|del} \\
P(\text{synapse change after } \Delta t) &= 1 - P(\text{synapse not changed})^{\frac{\Delta t}{t^{base}}} \\
p^{change}(\Delta t) &= 1 - \left(1 - p^{build|del}\right)^{\frac{\Delta t}{t^{base}}}
\end{aligned} \tag{4.6}
$$

4.3 Measuring neuron and network properties

In the following section we will demonstrate that our implementation can balance the neuronal responses and improves the information encoding in deeper layers. Further, we analyze the changes from the initial connectivity. To address the reduced dependency on initial values, we vary the initial state and compare the evolving structures. Finally, we control the stability of the mechanism in terms of preserving the retinotop organization and network convergence. We will focus for the evaluations on the first network layer, V1-layer 4, because of its well analyzable receptive field structures.

4.3.1 Spatial arrangement of formation and removal probabilities

Non-existing synapses surrounded by strong synapses get high formation probabilities, as the normalized weights of the surrounding synapses are accumulated and determine this probability. Whereas, with increasing distance to the receptive field center the formation probability decreases. This is because typical receptive fields have a spatially compact form with weaker synapses at their borders. Thus, with increasing distance from the receptive field center less synapses are within the neighborhood distance and their weights are weaker, which causes low formation probabilities. Note that existing synapses underly synaptic plasticity which shapes the receptive field. Which means for V1 simple-cells, synapses at the border of the receptive field have weak weights (because of the Gaussian envelop of the Gabor) and, thus, high removal probabilities (Fig. 4.3). In conclusion, our mechanism of synapse removal is shrinking the connectivity matrix to the size of the receptive field. Further, synapse formation allows gradual remodeling of the receptive field based on the learnings. This is achieved by randomly adding synaptic connections in the near surround. If these synapses gather strong weights by learning, the likelihood of forming more synapses in their neighborhood is increased, which is changing slightly

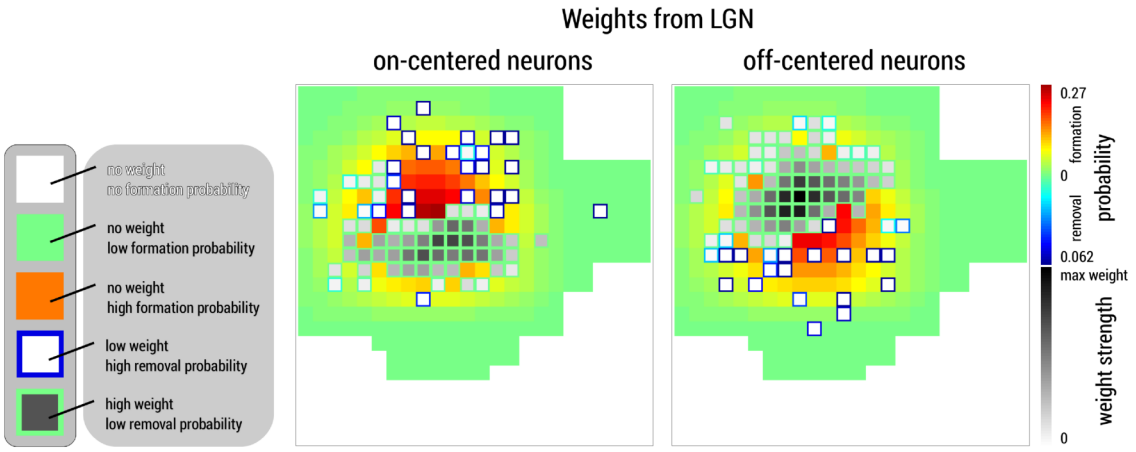

FIGURE 4.3: Spatial arrangement of synapse formation and removal probabilities. Gray boxes denote the connection strengths from on- or off-centered LGN neurons to a V1-L4 neuron. In the neighborhood of these connections (distance three) solid colored boxes indicate the probability that a synapse is formed. Potential synapses surrounded by strong synapses in both planes have the highest formation probabilities, whereas with increasing distance to strong synapses the formation probability decreases. The color of framed boxes indicate the probability for synapse removal. Mostly, synapses at the border of the receptive field have weak weights and high removal probabilities. Synapses are likely to be formed in the on-plane close to the center of the off-plane, these synapses are mostly removed as synaptic plasticity leads to low synaptic weights.

the receptive field shape. Formed synapses which do not develop sufficient high weight strengths are removed over time and have low probabilities forming more synapses in their surround.

4.3.2 Changes in the synapse amount

The synapse amount of the network should develop through structural plasticity to the amount needed for the individual functioning of each neuron. Further, structural plasticity should overcome insufficient initial settings by the modeler.

We measured the amount of synapses before and after learning (Tab. 4.1). To address the stochastic character of the structural plasticity and the dependency on the initial weights for synaptic plasticity, we ran 10 networks with randomly initialized weight values. We obtained minor differences between the different runs. The standard deviation (SD) across the 10 model runs for the total amount of synapses was 0.5%. We found that after training the network has 26.6% synapses less than initially. The most synapses are removed at the connections from LGN to V1-L4, about the half of the connections to the excitatory

neurons and about 60% to the inhibitory neurons have been removed. This was expected because of the character of the input, which consists of On- and Off-neuron responses. This neurons have just an activity when the opposing neuron has no activity, thus, no weights will be learned to both neurons at one position. In consequence, the half of the initial weights, which are emanating from both parts, are expected to be removed (cf. Fig. 4.3). When regarding the feedforward path in the network the other connections are more stable, e.g. from the excitatory neurons in V1-L4 to the excitatory in V1-L2/3 (-0.2%). Whereas the most connections decrease in their amount, several inhibitory connections increase, e.g. V1-L2/3 to V1-L2/3, V2-L4 to V2-L4, or V2-L4 to V2-L2/3. This might be caused by the initial definition of the connections, the modelers decision. All these connections have initially a low amount of synapses, also in comparison to the parallel excitatory pathway. In general it can be said that the connectivity in lower layers is less changed than in deeper layers. In V2 the most connections underly a change of $\pm$50% and more. However, the knowledge on the connection structure, choosing the right sizes of the connection patterns, and the more discrete steps of the sizes (2x2, 3x3), makes it difficult to chose optimal values without biasing the result by being too restrictive.

To get a better insight how the synapses are remodeled over time, we visualized the temporal development of the connections from LGN to the excitatory neurons in V1-L4 (Fig. 4.4). In the first period of learning, where V1-L4 underlies the strongest plasticity (see Sec. 4.3.5), the synapse amount increases to 220000 synapses. Than, the synapse amount decreases until it reaches a stable state around the 100000th stimulus presentation. The time course of the development of the other connections was comparable, but the phases of strong synapse changes are later dependent on layer depth and connection type (inhibitory was later than excitatory). Typically a stable amount of synapses was reached between the 100000th and 200000th stimulus presentation.

4.3.3 Development of the receptive field size under different initial conditions

To illustrate the influence of the modelers decision about connection sizes on the learnings, we compare four models where each V1-L4 neuron get inputs from 6x6, 9x9, 12x12, or 15x15 LGN neurons (connections are made to all neurons in the z-dimension). The models are trained on natural scenes (see Sec. 3), once without using structural plasticity and once

Target	Type	Source	Type	Initial	Final	SD	Change
V1-L4	Exc	LGN	Exc	194688	98087.3	±650.3 (0.7%)	-49.6%
		V1-L4	Inh	51076	55420.9	±363.4 (0.7%)	+8.5%
	Inh	LGN	Exc	48672	19108.5	±213.4 (1.11%)	-60.7%
		V1-L4	Exc	51076	12803.3	±228.7 (1.79%)	-74.9%
		V1-L2/3	Exc	28812	17956.6	±278.6 (1.55%)	-37.7%
		V1-L4	Inh	12600	12114	±176.3 (1.46%)	-3.9%
V1-L2/3	Exc	V1-L4	Exc	115248	114998	±997.1 (0.87%)	-0.2%
		V1-L4	Inh	46128	44059.6	±395.3 (0.9%)	-4.5%
		V1-L2/3	Inh	30276	38514.7	±592.5 (1.54%)	+27.2%
	Inh	V1-L2/3	Exc	30276	27658.6	±282.1 (1.02%)	-8.6%
		V1-L4	Inh	11532	10137.1	±158.8 (1.57%)	-12.1%
		V1-L2/3	Inh	7422	8486.4	±184.1 (2.17%)	+14.3%
V2-L4	Exc	V1-L2/3	Exc	61440	56191.6	±623.6 (1.11%)	-8.5%
		V2-L4	Inh	10000	15190.7	±171.6 (1.13%)	+51.9%
	Inh	V1-L2/3	Exc	15360	10797.7	±220.9 (2.05%)	-29.7%
		V2-L4	Exc	10000	6541	±129.1 (1.97%)	-34.6%
		V2-L2/3	Exc	18375	5711.9	±167.8 (2.94%)	-68.9%
		V2-L4	Inh	2420	3487.2	±45.9 (1.32%)	+44.1%
V2-L2/3	Exc	V2-L4	Exc	54000	44150.8	±782.6 (1.77%)	-18.2%
		V2-L4	Inh	9375	13727.5	±190.6 (1.39%)	+46.4%
		V2-L2/3	Inh	25500	7080.8	±561.3 (7.9%)	-72.2%
	Inh	V2-L2/3	Exc	25500	10148	±435.7 (4.29%)	-60.2%
		V2-L4	Inh	1700	3637.5	±126.1 (3.47%)	+114%
		V2-L2/3	Inh	7140	1122.8	±128.2 (11.4%)	-84.3%
Total				868616	637132.5	±3311.6 (0.5%)	-26.6%

TABLE 4.1: Change in the synapse amount. Overview on the initial synapse amount ("Initial") and the final amount after learning ("Final"). Sorted by the neuron type, target layer and source layer. The connection type is identical with the neuron type of the source layer. The initial amount of synapses is defined by the modeler. The final amount is averaged across 10 individual model runs using random weight values. The standard deviation (SD) of the 10 versions is given in synapses and percent. Further, the change between initial an final state is given in percent. Additionally, the change for the total amount of synapses in the network is given. Note, just learning synapses are considered.

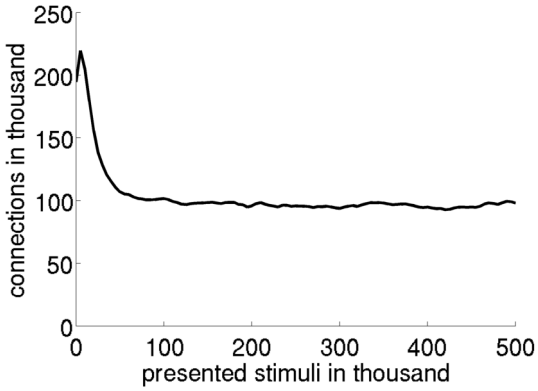

FIGURE 4.4: Development of the synapse amount between LGN and the excitatory neurons in V1-L4 over time. The plot shows the amount of the synapses for each 5000 stimuli presentations during the network training, averaged across 10 model runs. The initial synapse amount decreases after a short peak to approximately the half amount and remains than stable.

with structural plasticity.

Shape of the receptive fields

From each configuration we selected four neurons and visualized their receptive fields for illustration (Fig. 4.5; for details about the visualization see Sec. B.1). When structural plasticity was turned off, the restrictions for learning became very visible in the model using a 6x6 connection patch. The receptive fields span over the full connection matrix and clearly differ in size from the other models. Whereas, the neurons of the model with 15x15 connection patch developed larger receptive fields, while small receptive fields are also learned. Regarding the receptive fields of the different models in each model small receptive fields are learned. With increasing connection size a fraction of neurons learn larger receptive fields spanning up to the full size of the connection matrix.

When structural plasticity is activated, each model develops a comparable distribution of receptive field sizes. Also the most restricted 6x6 version, where the receptive fields of the version without structural plasticity seem visually to differ from the other versions, develops comparable receptive fields in appearance and size.

Spatial extents from Gabor fit

To obtain a more detailed insight on the distribution of receptive field sizes of the different models, we fitted the receptive fields of each excitatory neuron in V1-L4 to a Gabor function (see Sec. B.2) and compared their spacial extent, given by the Gaussian extents σ_x and σ_y. The distribution of the σ values confirmed the visual impression that the models, using structural plasticity, develop similar distributions of receptive field sizes indepen-

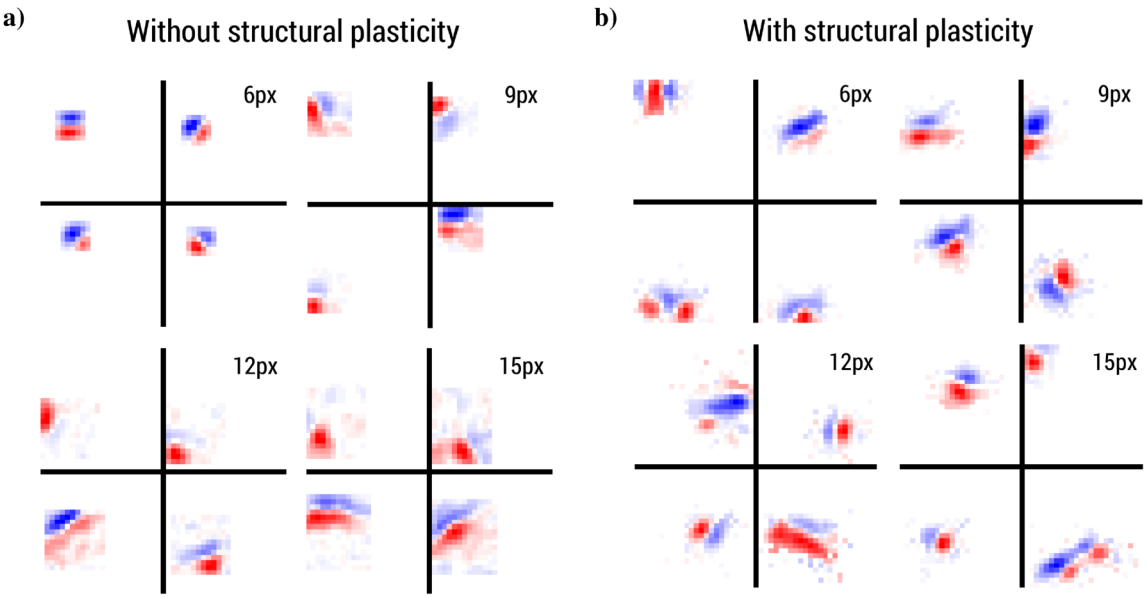

FIGURE 4.5: Receptive fields learned under different starting conditions. The initial connection matrix of each V1-L4 neuron was restricted to a patch of 6x6, 9x9, 12x12, or 15x15 LGN neurons. **a)** Four example receptive fields learned without structural plasticity. **b)** Learned with structural plasticity. With structural plasticity the neurons overcome the initial limitation and develop under each condition a huge, but similar, variety of receptive field sizes. The intensity of blue denotes the weight strength to off-center LGN neurons and the intensity of red to on-center LGN neurons.

dent from the starting conditions (Fig. 4.6). Whereas the Gaussian extents for the models without structural plasticity are narrow distributed below a value of two and spread with increasing connection matrix more and more to higher values. The distribution of the models with structural plasticity developed similar to the restricted model with the largest connection matrix of 15x15.

When regarding the distribution of the Gaussian extents, it becomes clear that the receptive fields of the models without structural plasticity are hindered to grow larger than the definition of their connectivity allows (Fig. 4.7a). Neurons in the model without structural plasticity (green dots) with the most restricted receptive field size develop just small receptive fields, having no large asymmetries. Increasing the receptive field sizes (e.g. black circles, red squares), leads to more neurons developing receptive fields with large asymmetries. Regarding the model variations with structural plasticity, all models develop similar receptive field size distributions, with many neurons with large asymmetries (Fig. 4.7b). Thus, similar to the histograms of the single variables the relation between width

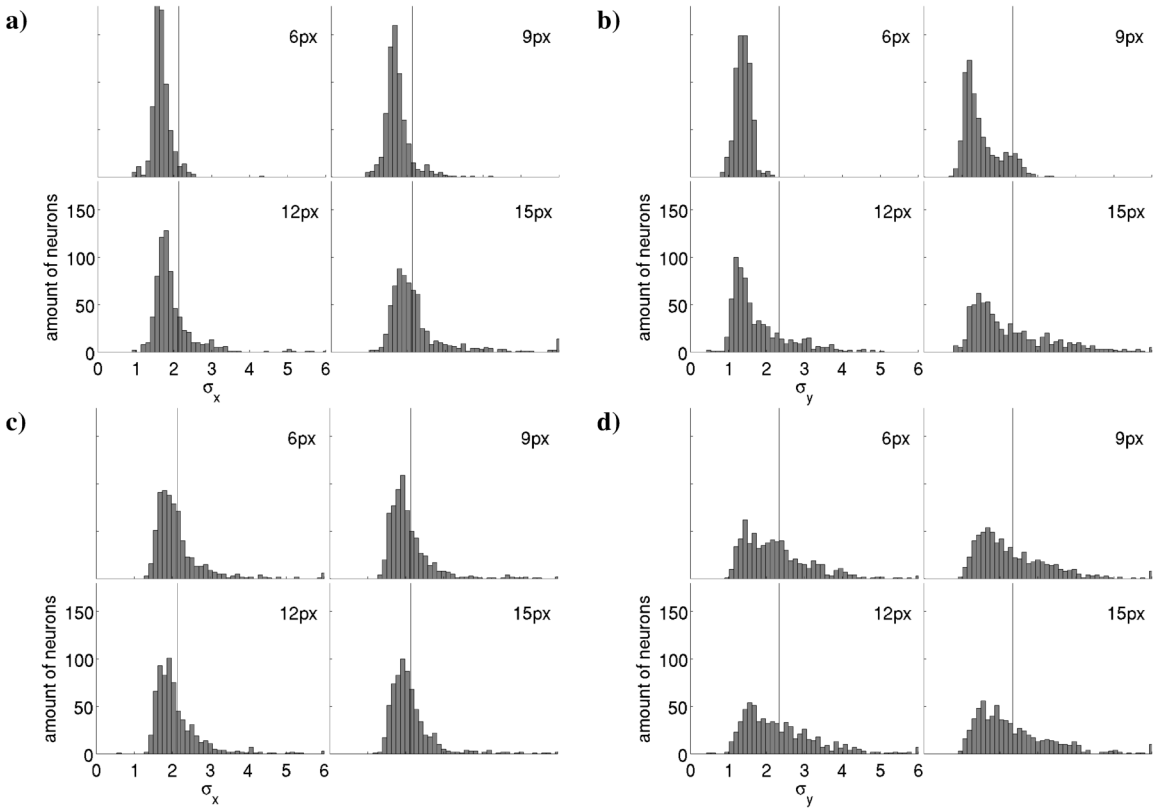

FIGURE 4.6: Histogram of receptive field extents, obtained with and without structural plasticity. The Gaussian extents σ_x and σ_y are shown, obtained by fitting the excitatory neurons in V1-L4 to Gabor functions. The vertical line indicates the mean of σ_x or σ_y over all model variations using structural plasticity. The axes are limited to $\sigma = 6$. The top row shows the model without structural plasticity **(a, b)** with the four different initial conditions, **a)** σ_x and **b)** σ_y. The bottom row shows the model with structural plasticity with the four different initial conditions, **c)** σ_x and **d)** σ_y. Whereas the model with structural plasticity **(c, d)** has similar distributions of the σ values under all conditions, the bias from the size of the connection matrix is clearly visible in a right shift of the distribution for smaller sizes.

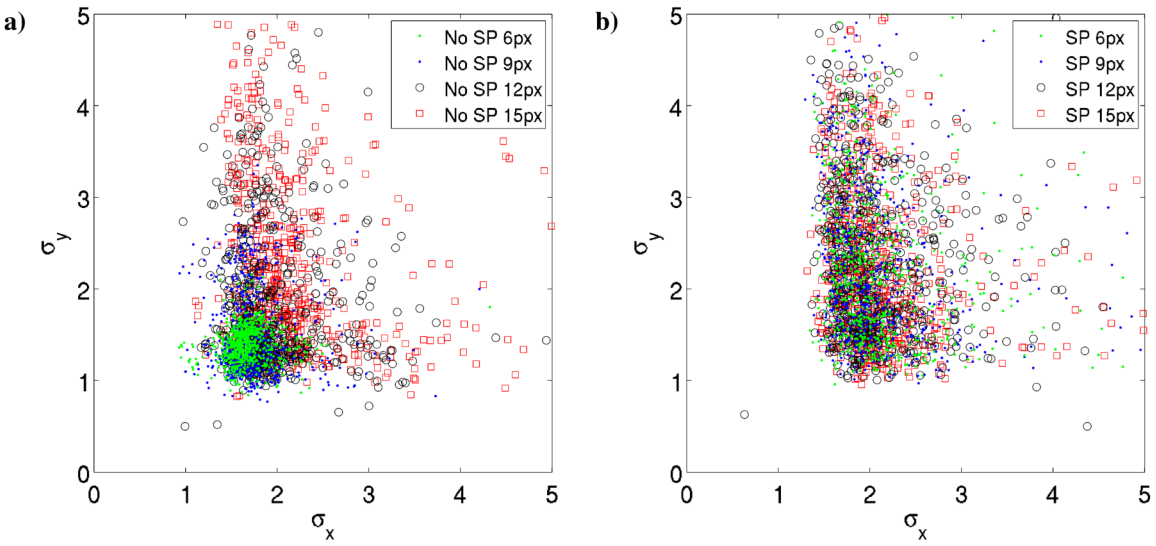

FIGURE 4.7: Relation of the receptive field extents, obtained with and without structural plasticity. The Gaussian extents σ_x and σ_y, obtained by fitting the excitatory neurons in V1-L4 to Gabor functions, are shown. **a)** The extents for the four different model conditions without structural plasticity. **b)** The extents with structural plasticity. With structural plasticity all models develop a similar $x-y$ relation. Without structural plasticity more restricted models lack neurons with strong asymmetries in their receptive field extents.

and height depends on the initial conditions. That means, allowing larger receptive fields leads to more neurons having an asymmetric relation between their receptive field extents.

To show, finally, that the receptive field sizes depend on the initial conditions of the model and that structural plasticity can overcome this initial bias, we compared the mean and median of the Gaussian extents between the different versions (Tab. 4.2). The mean of σ_x for the models with structural plasticity fluctuated between 2.11 and 2.17 for the different initial conditions. Whereas the mean extent for the models without structural plasticity strongly depended on the connection restrictions. It monotonically increases from 1.68 (6x6) to 2.3 (15x15). Similarly, the mean of σ_y, varies between 2.28 and 2.38 (with structural plasticity) and monotonically increases from 1.39 to 2.28 (without structural plasticity). A similar behavior is shown by the median values.

Because of the high computational expense of a single simulation, we did multiple model runs, 10 runs each, for just two model versions. We measured the standard deviation of the mean and median values of the neurons Gaussian extents. We chose the model version with 12x12 initial receptive fields, with and without structural plasticity, because this configuration was the one we use for the most other evaluations (Tab. 4.3). The stan-

Inital RF size	Condition	mean σ_x	mean σ_y	median σ_x	median σ_y
6px	SP	2.1686	2.2834	1.9705	2.0945
	NoSP	1.6794	1.3895	1.6515	1.3869
9px	SP	2.1143	2.3388	1.9430	2.0438
	NoSP	1.7418	1.5066	1.6863	1.3326
12px	SP	2.1223	2.3766	1.9369	2.1399
	NoSP	1.9622	1.7815	1.8193	1.4908
15px	SP	2.1362	2.3605	1.9857	2.0618
	NoSP	2.2999	2.2812	2.0081	1.8884

TABLE 4.2: Mean and median extents, obtained with and without structural plasticity. The Gaussian extents are obtained via Gabor fitting of the receptive fields. The neurons from models with structural plasticity developed similar receptive fields regardless of the initial receptive field size. Whereas the neurons from models without structural plasticity showed a direct relation between size of their connection matrix and their finally learned receptive field size.

Inital RF size	Condition	mean σ_x	mean σ_y
12px	SP	2.1208 ± 0.0276	2.3638 ± 0.0376
	NoSP	1.9980 ± 0.0250	1.8050 ± 0.0248
		median σ_x	median σ_y
12px	SP	1.9563 ± 0.0126	2.0750 ± 0.0283
	NoSP	1.8179 ± 0.0076	1.5077 ± 0.0102

TABLE 4.3: Variation of the mean and median extents, obtained with and without structural plasticity. We compared the mean and median Gaussian extents of 10 model runs for each of the 12x12 models and report their standard deviation. The largest observed standard deviation was 1.59%.

dard deviation of the obtained average extents was below two percent for all conditions, which is sufficiently low to rely on the results from the other single model runs.

4.3.4 Structural plasticity does not harm retinotop organization

The model's structural plasticity mechanism enables the neurons to change their receptive field size by a stochastic process of synaptic growth and retraction. This process relies only on the synaptic weight strengths in the neighborhood, regardless of their position within the layer. Thus, the receptive field development is not biased to a certain position, except by its initial placement. This is contrary to the cortex, where neurons and their parts are embedded in the very dense cortical matter which impedes extensive connectivity changes

and the development of very large dendritic arbors of a substantial amount of neurons.

To control that our unrestricted approach does not harm the retinotopic organization of the model neurons, we compared the learned connectivity to the initial connections. We did this just for layer V1-L4 where the receptive fields are sufficiently smaller than the input layer size and, thus, a position is comparable. We compared the initial connectivity to the final connection matrix. We just regard the existence of a connection without regarding the weight strength, this means an existing weight can also be zero or very weak and, hence, has negligible influence on the neuron's behavior. For visual comparison we plotted the initial receptive field and the final connectivity together. The connection onto neurons in the depth plane of the cortical grid is collapsed. This means that a connection is drawn if any weight to a neuron with the x-y coordinates exists. We visualized the connections from LGN to the excitatory neurons in V1-L4 and the connections from these to the inhibitory neurons of the same layer. Further, we used the x-y centers from the Gabor fit (Sec. B.2) of the weight matrix of each neuron as measure for the receptive field position. We compared the initial position of their receptive fields with the mean of the fit centers of all excitatory V1-L4 neurons at the same x-y grid position.

Figure 4.8 shows the initial connection matrices and the connectivity after learning of the first 100 excitatory neurons of V1-L4 (Fig. 4.8a), for comparison we show the learned weights besides (Fig. 4.8b). Just a few neurons developed receptive fields (bright dots) far away from their initial weights (dark squares). Similar results are shown for the excitatory connections from the excitatory to the inhibitory neurons of V1-L4 (Fig. 4.9). The most neurons have their connection center within the area of the initial connections. However, connections from the excitatory neurons to the inhibitory appear less localized, this might has its reason in the larger initial receptive fields which can span up to 11 by 11 out of 13 by 13 positions in the cortical grid. Inhibitory neurons are initially connected to receive connections from all excitatory neurons which could have an sufficiently overlapping receptive field. Backward connections, from inhibitory to excitatory neurons, are even less specific when regarding just the existence. This is because the inhibitory learning rule leads to positive weights also when neurons do not correlate.

To gain further insights whether the retinotop organization is preserved, we compared the centers of the receptive fields of the excitatory neurons, obtained by the Gabor fit, with their position in the cortical grid. When the retinotop organization is preserved, the neurons should show a strong relation between their cortical position, from which we

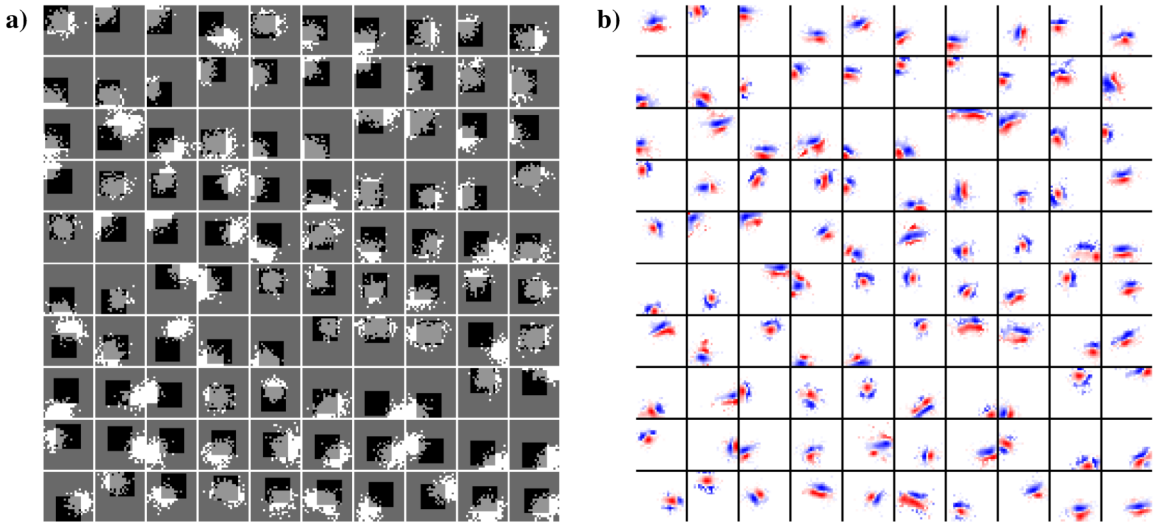

FIGURE 4.8: Connection matrices from LGN to V1-L4 excitatory neurons before and after learning. Each tile shows the formed connections of an individual neuron, data for the first 100 V1-L4 excitatory neurons are shown. **a)** The connections before (dark squares) and after learning (bright dots) are shown. **b)** For comparison, the weight matrix of each neuron is individually normalized to use the full color range.

FIGURE 4.9: Connection matrices from excitatory to inhibitory neurons in V1-L4 before and after learning. Each tile shows the formed connections of an individual neuron, data for the first 100 inhibitory neurons are shown. The connections before (dark squares) and after learning (bright dots) are shown.

determine their initial retinotop connectivity, and the receptive field center after learning. We plotted the average x- and y-center of the fit in relation to the position in the cortical

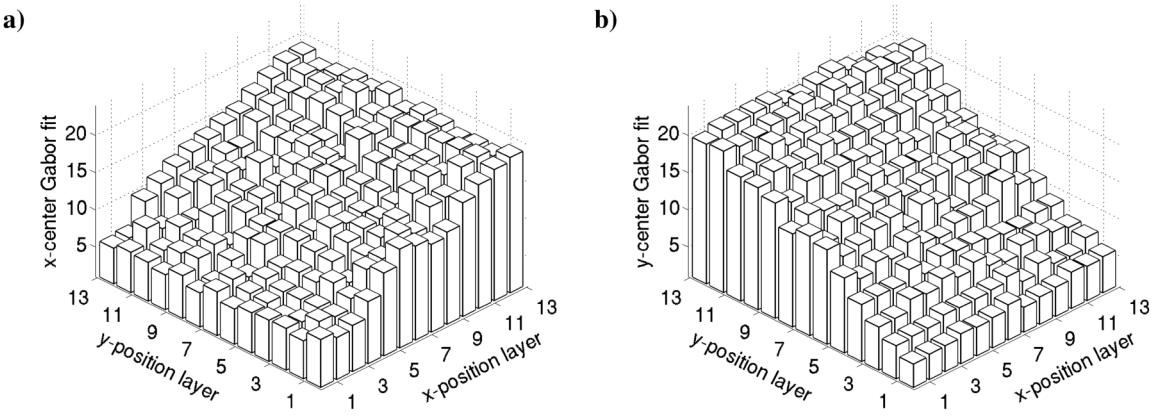

FIGURE 4.10: Relation of the V1-L4 excitatory neuron's cortical position to their receptive field position. Each bar indicates the average of all neurons at a one position on the cortical grid over the neuron's receptive field position in **a)** x-dimension and **b)** y-dimension in the image space.. The receptive field centers are determined through Gabor fitting the connection matrices. The receptive field position shifts analogue to the cortical position of the neurons.

grid. We found that the fit's x-center increases with the cortical x-position, regardless of the cortical y-position, analogue the y-center increases with the y-position in the grid (Fig. 4.10).

With that, we could visually show that the developed connection matrices are close the initial matrices. When we regard the position of a neuron in the cortical grid, in comparison to the position of its receptive field in the visual domain, we found that the retinotop organization was preserved. That means, the receptive fields are ordered in the cortical grid as in the visual domain.

4.3.5 Stability of structural plasticity

An additional source of plasticity within a model can make it more robust or less stable. The stochastic character of structural plasticity can reduce the stability of the learnings by its induced random fluctuations. We evaluated the amount of weight change during learning as measurement of stability and compared the model version with structural plasticity to the configuration without.

To measure stable learning we regarded the stability of the feedforward connections. These connections mainly determine the behavior of the neurons and should underly rapid changes when the response character of the neuron change. Therefore, we recorded the connection weights from LGN to V1-L4 every 5000 stimulus presentations during the

training. For each neuron we calculated the normalized mean square error (cf. (Spratling, 2012)) between its weight vector w_t at time t and its previous recording $w_{t'}$ (Eqn. 4.7). The resulting $NMSE_t$ value is averaged over all neurons to capture the overall change of the network. We calculated the average weight change for 10 independent model runs with random initial weights.

$$NMSE_t = \frac{\sum(w_{t'} - w_t)^2}{\sum w_(t)^2} \tag{4.7}$$

To get a measure for the stability of the model we additionally determined the rate of weight change, ignoring the initial phase of strong plasticity. Therefore, we calculated from the recorded time series of NMSE values the median value, which is insensitive to strong changes in short time periods and should capture the average change in the system. The average of this value over the 10 different model runs is used to compare the two different model versions.

We found that the model with structural plasticity had a comparable rate of change to the model without structural plasticity. After an initial phase of strong synaptic changes in the first 25 thousand training stimuli, both models remain at a low rate of change until the end of the training (Fig. 4.11). The average across multiple model runs of the median change value is 0.1251 (NMSE) for models without structural plasticity and 0.1191 with. In the initial phase the changes have been stronger in the models with structural plasticity, which presumably results from the additional degrees of freedom in the remodeling of the weight matrix. Thus, we conclude that structural plasticity does not harm the stability of the model.

4.4 Conclusion

We have shown that our structural plasticity mechanism modifies receptive fields mainly in its surround by random formation of new synapses close to existing synapses and removes weak synapses. The mechanism is build upon the principles that the weight strength is related to the synaptic lifetime and that the formation of new synapses takes place in vicinity to active ones. Our model is with and without structural plasticity able to learn various Gabor-like simple-cell receptive fields in V1-L4. However, without structural plasticity the development of the receptive field sizes depends on the limitation of the individual connection matrix of each neuron. Without, each of the four regarded model configu-

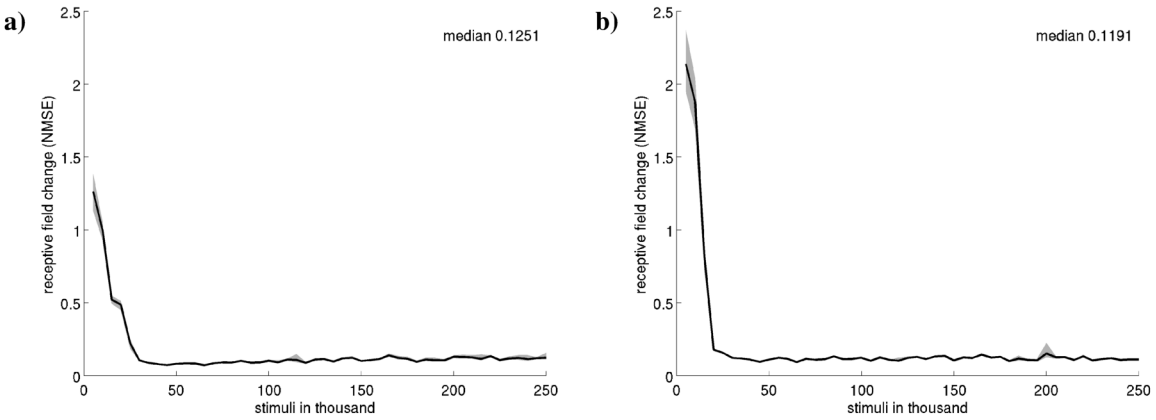

FIGURE 4.11: Change of the LGN to V1-L4 excitatory synapses over time. The normalized mean square error (NMSE) between the weight matrices after each 5000 stimulus presentations is shown. **a)** For the model without structural plasticity and **b)** with structural plasticity. The gray shaded area indicates the maximum and minimum change of 10 different model runs. Further, the median of the changes to 500 thousand stimuli over the multiple model runs is reported. The model with structural plasticity showed a higher rate of change in the beginning, but slightly lower rate of change in later phases of the learning.

ration develops a different distribution. We showed that structural plasticity overcomes this restriction and that similar receptive field size distributions can be developed independent from the initial connection matrix sizes. To enable neurons in the downstream layers to learn from spatially related features, a retinotop organization during the learning is required. We found that structural plasticity sufficiently preserves the initial retinotop organization of the neurons, despite we implemented no constraint to ensure this. Structural plasticity can be a source of additional random fluctuations in the learning process, thus, we controlled the network changes over time. We found a similar change rate after an initial strong plasticity phase in networks with and without structural plasticity. Further, we measured the change of the synapse amount over time and found a similar course of strong changes in the beginning and stability in later phases of the learning. We found the changes of the connectivity through structural plasticity meaningful. It removed unnecessary connections and formed new connections, where the connectivity was not suitable defined by the modeler. Thus, we can conclude that structural plasticity overcomes the modeler's bias of defining the network connectivity, while being sufficient stable. It does not harm the learning process. Our stochastic approach captures several ideas from neurophysiology and depicts a practically usable implementation within networks with a meaningful neighborhood like organization.

5 Intrinsic Plasticity

This chapter introduces the intrinsic plasticity mechanisms. First, we describe why we need this plasticity. We introduce the related computational models and their implementations. Then, we present our neuron model and the implementation of the plasticity mechanisms. Subsequently, we evaluate the functioning, stability, and effect on the activity distribution. Parts of the chapter have been published in Teichmann and Hamker (2015).

5.1 Introduction

One of the most important criterion of biologically plausible learning principles is their locality, i.e. a neuron has only access to its own state and incoming signals. Common formulations of Hebbian learning use only information local to synapse or neuron. Indeed, such formulations can not ensure that a population of so defined neurons will learn a codebook representing the input's manifold. By extending such populations by plausible mechanisms of inter neuron interaction, as inhibition, it can be shown that an adequate representation of the input can be learned (e.g. Földiák, 1990; Wiltschut and Hamker, 2009). A common plausible principle to learn such connections is anti-Hebbian learning (e.g. Földiák, 1990; Wiltschut and Hamker, 2009; Teichmann et al., 2012). However, this principle aims only to reduce correlations between neurons based on their coactivity. It has no objective ensuring an adequate input encoding by a population of neurons. This means that information encoding can be just optimal when the neuron responses are independent, i.e. no information is duplicated, and all information in the input lead to a response (Simoncelli and Olshausen, 2001). However, Hebbian learning depends on the association between pre- and postsynaptic activity. If a presynaptic neuron shows, by average, more activity, its connection will be more strengthen. Hence, it is very likely for more complex input data, or in higher layers of abstractions, that some neurons will show higher activity for the pattern they encode than others. Which in turn will enforce the imbalance in subsequent layers (cf. Diehl and Cook, 2015). This means, the information, which neurons with

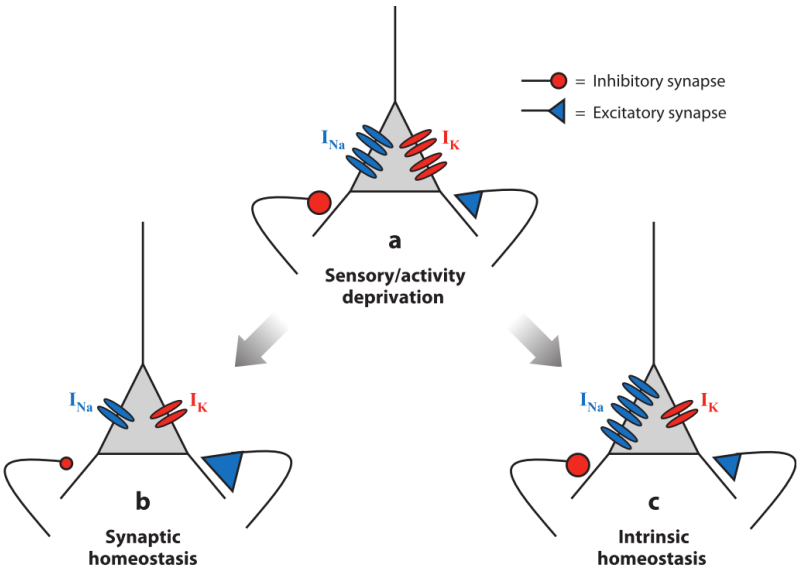

FIGURE 5.1: "Two fundamentally different mechanisms for the homeostatic regulation of neuronal firing. (a) Neuronal activity is determined both by the strength of excitatory and inhibitory synaptic inputs and by the balance of inward and outward voltage-dependent conductances that regulate intrinsic excitability, here illustrated as the relative number of Na (blue) and K (red) channels. Neurons can compensate for reduced sensory drive either by using synaptic mechanisms to modify the balance between excitatory and inhibitory inputs (b) or by using intrinsic mechanisms to modify the balance of inward and outward voltage-dependent currents (c)." Figure and description taken from Turrigiano (2011).

higher average activities convey, will dominate the information the efferent neurons will learn to encode.

The human neural system has various mechanisms to stabilize its functioning (Turrigiano and Nelson, 2004; Turrigiano, 2011). Beside homo- and hetero-synaptic regulations of the weight development, as synaptic scaling, it has been found that neurons preserve their average firing rate over time, independent from their synaptic strengths (Zhang and Linden, 2003). If a neuron is highly stimulated, or weakly, over a long period of time it can be observed that its sensitivity to this stimulation decreases, or increases, and the neuron returns to its previous activity regime (Fig. 5.1). This means, its intrinsic excitability is adapted, called intrinsic plasticity (Zhang and Linden, 2003; Turrigiano, 2011).

For first computational implementations of intrinsic plasticity it has been speculated that neurons try to approach an exponential firing regime (Stemmler and Koch, 1999; Triesch, 2005a,b). This is because exponential distributions are found efficient for information encoding, for a fixed average activity in a single neuron case (Simoncelli and Olshausen,

2001), as well as transmitting information energy efficient (Rolls and Treves, 2011; Levy and Baxter, 1996). Stemmler and Koch (1999) modified a Hodgkin-Huxley model by voltage dependent conductances. The conductances are increased or decreased by a constant value when the average voltage in the regarded compartment is above or below a certain value. With that, they regulated the mean activity of the neuron. They found that the model changed the activation function so that the responses evoked by a Gaussian input distribution become exponentially distributed. Triesch (2005b) introduced a sigmoidal activation function, where the baseline and slope of the inward current can be configured. These values are adapted during learning to match the mean and the variance of a desired exponential distribution. Triesch (2005a) introduced online minimizing the Kullback-Leibler divergence of the neuron activity to an exponential function as alternative approach to determine parameters for the slope and baseline value. Later Savin et al. (2010) demonstrated this approach in a stochastically spiking model, learning independent components from natural scenes. These approaches explicitly modify the neuron's activation function to achieve exponential firing. However, Triesch (2005a, 2007) have been criticized that, their approach to enforce an exponential distribution, drives the activation function to biologically implausible high values and that enforcing sparseness rather than an exponential distribution would be sufficient (Elliott, 2014). Altogether, it is still unclear whether the objective of cortical neurons is an exponential regime or how they achieve it.

We speculate that the aspect of the exponential regime is not an objective of cortical neurons, rather it is a byproduct of the neural circuit developing sparse representations. Instead of enforcing an exponential activity distribution, we hypothesize the intrinsic plasticity mechanism aims to stabilize the operating point of the neurons so that the brain is not wasting resources for non responding cells or hyperactive ones. Thus, we aim to control the two most important moments of neural activity, the mean and the variance, by adapting the slope and the threshold of the rectified linear activation function of our model neurons (Sec. 5.3). Beside sigmoidal activation functions, rectified linear activation functions have been very common in the past and recently in deep neural networks. There is also biological evidence that this function type is an adequate description for cortical neurons Ringach and Malone (2007). However, this function is mainly linear, except its rectification. Hence, mathematically it can not transfer any input distribution into an exponential one. We implemented a modifiable rectified linear activation function and the control of its parameters and demonstrate first its effectiveness for stabilizing the neural response prop-

erties (Sec. 5.4). Than, we analyze the information encoding in deeper network layers. We compare different parameterizations of the plasticity rule. Finally, we examine if the activities of our model neurons are exponentially distributed and whether this is caused by our intrinsic plasticity mechanism.

5.2 Neuron model with intrinsic parameters

We have modified the rectified linear neuron model of Wiltschut and Hamker (2009); Teichmann et al. (2012) with parameters controlling the intrinsic excitability. With this additional parameters, we aim to achieve that all neurons are equally participating in the encoding of their input (Teichmann and Hamker, 2015). Weakly active neurons should be strengthened to facilitate the learning of unrepresented patterns and strong neurons should be weakened to prevent that few units dominate the output. This is done by regulating the first moments of activity of a neuron, i.e. the mean and variance. We employ a rectified linear activation function, where the parameter θ controls the mean activity by increasing or decreasing the activity threshold, which in fact shifts the mean. The slope a modulates the response strength for inputs exceeding the activity threshold.

We define the change of the membrane potential m_j of a neuron j (Eqn. 5.1) by its excitation, reduced by its inhibition, and further reduced by its threshold θ_j. The resulting value is weighted by the factor a_j, defining the slope. To obtain the membrane potential change, we subtract the current membrane potential. The excitation and inhibition (neural drive) are calculated as weighted sum. The firing rate r_i denotes the activity of an afferent excitatory neuron i and w_{ij} is the weight between the two neurons. Similarly, r_k denotes the activity of an afferent inhibitory neuron k and c_{kj} the weight.

$$\tau_m \frac{dm_j}{dt} = \underbrace{a_j}_{\text{slope}} \cdot \Bigg(\underbrace{\sum_i w_{ij} r_i}_{\text{excitation}} - \underbrace{\sum_k c_{kj} r_k}_{\text{inhibition}} - \underbrace{\theta_j}_{\text{threshold}} \Bigg) - m_j \tag{5.1}$$

After calculating the membrane potential, we transfer this activity into the output firing rate r_j of the neuron j. Therefore, we rectify the membrane potential, where $(.)^+$ denotes the clipping of values below zero. Values above one are transfered nonlinear, using an

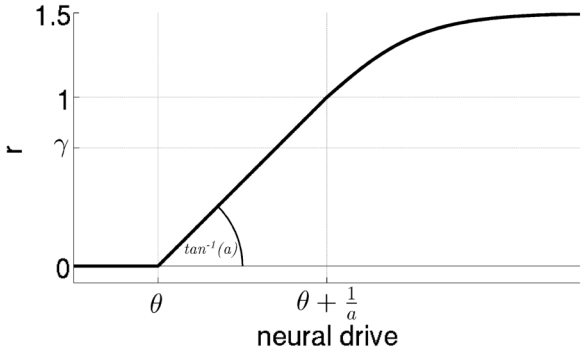

FIGURE 5.2: Activation function with intrinsic parameters. The neuron response for neural drives above θ grow linearly, until an activity of 1. Above 1 the activity saturates to its limit of 1.5. The slope is defined by the parameter a. Above γ the homeostatic regulation of our learning rule is adapted to impede further weight increase.

exponential saturation function.

$$r_j = \begin{cases} 0.5 + \frac{1}{1+e^{-3.5(m_j-1)}} & \text{if } m_j > 1 \quad \}\text{saturation} \\ (m_j)^+ & \text{else} \quad \}\text{rectification} \end{cases} \tag{5.2}$$

The resulting course of the activation function is illustrated in Fig. 5.2. If the neural drive exceeds the threshold θ of a neuron its firing rate r increases linearly, until its maximal "normal" activity of 1 is reached. Above 1 the activity saturates to its limit of 1.5. Note, neurons rarely reach the saturation zone as the synaptic learning reduce the synaptic weights for high activities. The used initial values can be found in Table A.3.

5.3 Intrinsic plasticity mechanisms

We aim an equal participation of the neurons in encoding the inputs by controlling their first activity moments (cf. Triesch, 2005b). This are, the mean and variance of their activity. Therefore, we adapt the parameters θ_j and a_j for each individual neuron so that an activity regime with mean θ_{target} and squared activity a_{target} is approached. A small constant drift is added to the adaption of the parameters, giving the neurons a small bias to prefer a minimal modified activation function. Further, this prevents neurons in very deep layers, which learn long periods from very noisy input, from developing extreme values. In the following the mechanisms for threshold adaption, slope adaption, and drift are described. The used parameters can be found in Table A.4.

5.3.1 Threshold adaption

The input currents to a neuron (neural drive) have to breach a positive threshold value when a neuron should become active. However, also negative values are possible, than the threshold acts as a baseline activity and keeps the neuron active, also when excitation and inhibition would extinguish each other. The threshold value is comparable to the bias value, used in the multilayer perceptron. The adaption of the threshold parameter θ_j of neuron j is realized by increasing the threshold when the neuron is more active than the target value θ_{target} and decreasing it when it is less active (Eqn. 5.3).

$$\tau_\theta \frac{d\theta_j}{dt} = (r_j - \underbrace{\theta_{target}}_{\text{target mean activity}}) - \overbrace{\delta(\theta_j)}^{\text{drift}} \tag{5.3}$$

Consequently, it becomes more and more difficult for a neuron to fire when its activity was a long time above the target value, because its threshold increases. Vice versa, when a neuron, by average, is active below the target value its threshold decreases and firing becomes easier. When a neuron is just active below the target value or becomes inactive, than the threshold decreases to negative values and lifts the neuron to activity. Since a neuron can just learn when it is active, this mechanism revives silent neurons. Further, when a neuron became inactive, because of receiving strong inhibitory currents, the increased baseline can make it responding to under represented patterns, which where characterized by reduced network activity and, as consequence, reduced inhibition. By this, potentially a neuron can develop a high baseline activity and its activity is just modulated through inhibition, which rises when an input is represented by other neurons. To avoid permanent high threshold values, we added a small constant drift function $\delta(x)$ (Sec. 5.3.2) in the direction of the threshold's origin (typically zero). High values can be caused by stimulating the neurons with very noisy input, inducing low sparseness and strong inhibition. This can be the case in deep layers at the beginning of the network training, when starting from random uniform distributed weights. It would be also possible that an equilibrium is found at very high threshold values, despite that low threshold values would be possible for functioning. As target activity we use the mean activity of the neuron population. This has the computational advantage that the target value becomes stimulus contrast invariant and no individual parameter has to be determined for different populations, layers, neuron

types, or input datasets. The time constant τ_θ is set to a sufficient large value (10000ms) so that changes in the input and changes in the average input contrast just have a longterm effect. Further, it is chosen that learning does not totally change the system before adaption takes place.

5.3.2 Slope adaption

Additionally to the regulation of the threshold it can be important to increase, or decrease, the sensitivity of a neuron. Theoretically, an equilibrium state for the threshold can also be reached with a constant firing rate, which means that the neuron activity carries no stimulus information anymore. Whereas, when a neuron has an activity equal to the target activity, it is not possible, without fluctuations in the neuron's activity, to have a squared activity equal to the so called target squared activity - when not picking target value for the slope as square of the target value for the threshold ($a_{target} = \theta^2_{target}$). However, it would be anyway unlikely that the activity of a neuron starting with random weights degenerates to a constant. Consequently, we regulate the slope a_j of a neuron j (Eqn. 5.4), which modulates the strength of the neuron's response on an input. The slope decreases when the squared activity of the neuron is above the target activity a_{target} and increases when it is below. Here again, we use a constant drift to the origin of $a_j = 1$.

$$\tau_a \frac{da_j}{dt} = \underbrace{(a_{target} - r_j^2)}_{\text{target squared activity}} - \underbrace{\delta(a_j - 1)}_{\text{drift}} \tag{5.4}$$

5.3.3 Drift

We add a small constant drift $\delta(x)$ in the direction of the initial values $\theta_j = 0$ and $a_j = 1$ (Eqn. 5.5). This gives the neurons a small bias ε to prefer a minimal modified activation function. The drift is positive for the values $x > 0$, negative for $x < 0$, and zero for $x = 0$. Further, it prevents neurons in very deep layers, learning long periods from very noisy inputs, from developing extreme values.

$$\delta(x) = \varepsilon \cdot sgn(x) \tag{5.5}$$

5.4 Measuring neuron and network properties

In the following section we will demonstrate the functioning and stability of the proposed mechanisms. We show that the regulation of the mean and variance achieves its goals and that particularly deeper layers profit. Further, we examine the influence of the drift on the development of the parameters. Finally, we analyze the resulting activity distributions.

5.4.1 Development of threshold and slope during learning

The intrinsic regulations change the neurons activation function. The amount of the changes and the time course of the parameters are of particular interest to assess the functioning and stability of the regulation.

Therefore, we visualized the distribution of the threshold and slope parameter after each 5000 training stimuli over the full network training of 500000 stimulus presentations (Fig. 5.3, for further results see Appendix C.1). The obtained temporal histogram has 0.02 bins in the value domain, where the color of each bin indicate the amount of neurons having this threshold or slope value. The color range is chosen logarithmic to make values beside the origin more visible.

We see that the intrinsic parameters closely cluster around the origin of the parameters, i.e. around zero for the threshold θ and around one for the slope a. Just at the beginning of the network training, where the neuronal weights underly strong changes (see Sec. 4.3.5), a larger fraction of neurons develop parameters apart from the origin. The amount of values more than 0.1 apart from the origin was up to 50 percent for early phases of the network training, whereas in later training phases approximately 90 percent of the parameters have been in a range of 0.1 around the origin.

5.4.2 Effective modulation of mean activity and variance

The intrinsic plasticity is intended to regulate the mean response of each neuron and its variance with the goal to achieve equal values across all neurons within a population. This should reduce a bias for particular synaptic connections of the Hebbian learning rule. Such a bias would gather more and more resources in terms of high weight values onto few presynaptic neurons, this effect would increase over the hierarchy.

To investigate the effectiveness of the proposed intrinsic plasticity mechanism, we ob-

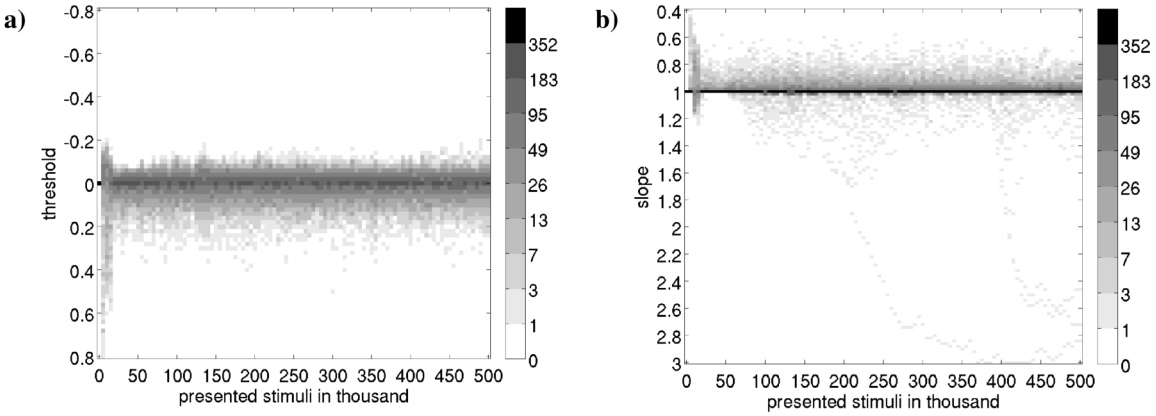

FIGURE 5.3: Development of the intrinsic regulation parameters over time. a) Shows the histogram of the threshold parameter θ for the excitatory neurons in V1-L4 at each 5000 stimulus presentations over the full network training and b) shows the slope parameter a. In the early phases with strong synaptic plasticity the intrinsic regulation is strong. In later phases of the network development the values cluster closely around the parameter's origin. The gray tone indicate the amount of neurons within a bin. The bin size is 0.02. Note, the color scale is logarithmic to improve the visibility of the parameter distribution.

tained the response statistic for each network neuron after learning. Therefore, we turned all plasticity mechanisms off and presented 100000 randomly selected natural scene patches to the network. At the end of each presentation period (100ms) we stored the activities. With that values, we calculated the mean and variance for each neuron from its 100000 responses. For comparison, we further repeated the same analysis for three model variations, where we: *1)* turned off the intrinsic plasticity, *2)* used only the regulation of the mean, and *3)* used only the regulation of the variance.

Effect with full intrinsic plasticity in comparison to no intrinsic plasticity

For the model with full intrinsic plasticity, we obtained in each neuron population a Gaussian distributed mean (Fig. 5.4ab) and variance (Fig. 5.4ef). The model without intrinsic plasticity the excitatory neurons of the V1-layer 4 seems to have a similar distribution (Fig. 5.4cg), however, the distribution is broader and a few neurons developed significantly higher mean and variance values. This results in a catastrophic effect in the deeper layers (Fig. 5.4dh), where the distribution changed so that a few neurons dominated the population code with their high mean activities and variances, whereas the majority of neurons had no or low activities. The catastrophic effect is presumably caused by the inhibitory mechanisms, which raises inhibition between all correlating neurons until they

learn something else or remain silent. Indeed, learning other patterns is tough when few presynaptic neurons have much higher activities, which causes that these connections are more strengthen than others.

Effect when regulating only one intrinsic parameter

It is obvious to regulate the variance of a neuron beside its mean. An equal mean can be achieved by just inducing a constant current, as possible with our threshold θ, but any information would be lost with zero variance. On the contrary, inhibitory plasticity is thought to achieve decorrelated activities of the neurons and so enforces them to have a variance in their firing. Consequently, we tested whether using just one regulation on the mean or the variance is effective to regulate mean and variance of the neurons. Therefore, we deactivated the plasticity of the other respective parameter and we applied the same testing procedure as before.

When we regulated just the threshold, we observed similar results (Fig. 5.5cd) as with full intrinsic plasticity (Fig. 5.5ab). The mean and variance became Gaussian distributed, no neurons had been inactive, and nearly no showed low variance. In deeper layers few units with higher mean and variances can be found (Fig. 5.6cd). Thus, holding the neurons activity within a certain range keeps them participating in the encoding of input patterns and the patterns learned through synaptic plasticity cause a similar variance of the activities.

When we regulated just the slope, the distribution of the neurons mean in the second layer became broader and several neurons showed a higher mean activity than the average (Fig. 5.5e). The variance was narrowly tuned, but several units had variances close to zero (Fig. 5.5f). When regarding even deeper layers also the regulation of just the slope seems to be effective to keep all neurons participating in the encoding of stimuli (Fig. 5.6ef). No neuron with zero mean and just very few neurons with very low variance have been found. However, while the distribution of the variance is narrow, the distribution of the mean became very broad. Nevertheless, the distribution strongly differs from the model without intrinsic plasticity, where the most neurons became inactive or showed very low responses (Fig. Fig. 5.5gh, 5.6gh).

Thus, we found that the regulation of one parameter is enough to achieve a vital encoding of stimuli also in deeper layers. As expected, when regulating the parameter for the mean or the variance the respective distribution becomes narrower. Nevertheless, the reg-

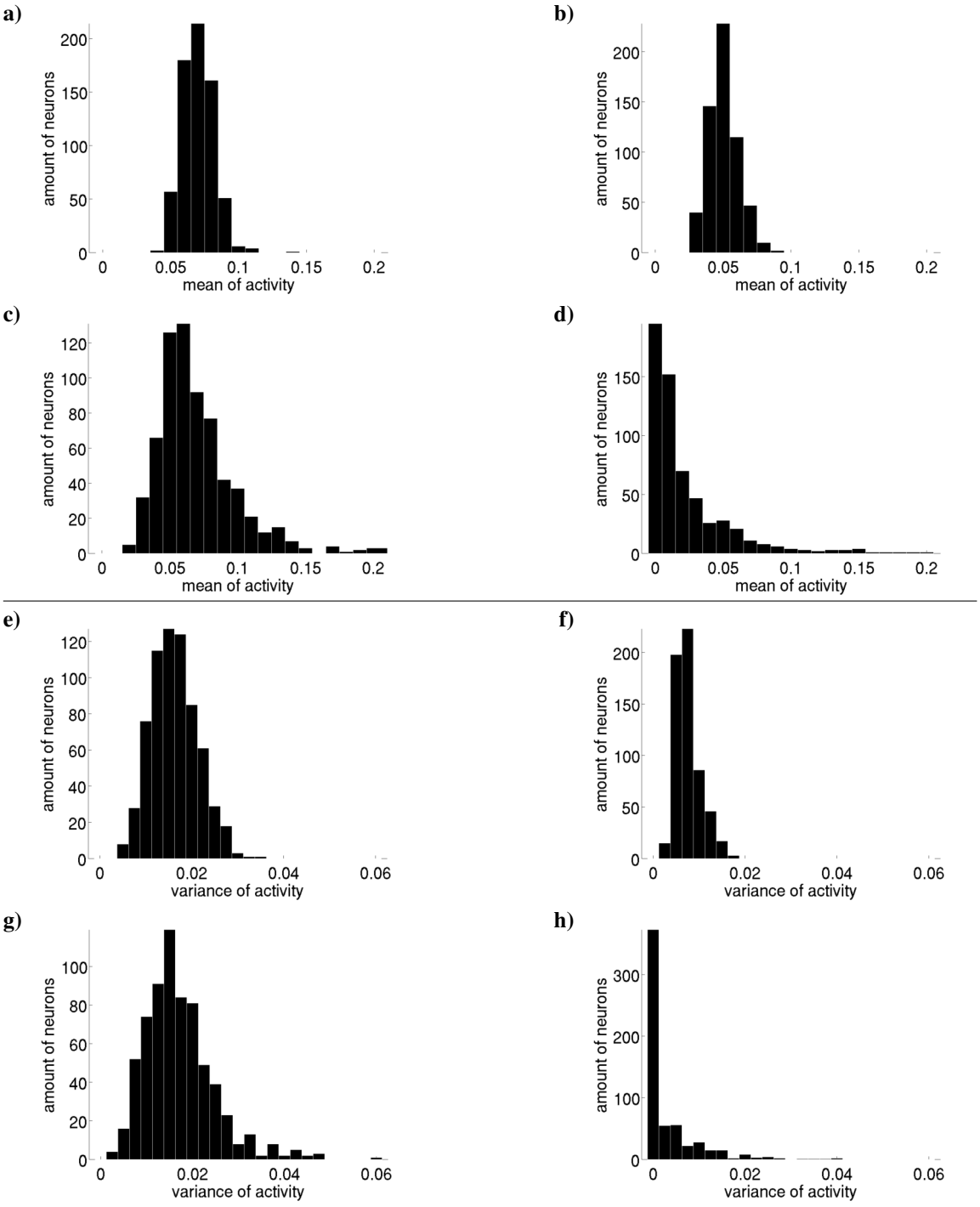

FIGURE 5.4: Histograms of the mean and variance of the neurons activity. The left column (a, c, e, g) shows results from excitatory V1-layer 4 neurons, the first layer, and the right from excitatory V1-layer 2/3 neurons, the second layer. The mean of activity (a, b) and the variance of activity (e, f) are obtained with full intrinsic plasticity. Whereas the subsequent row (c, d; g, h) shows the results obtained without intrinsic plasticity.

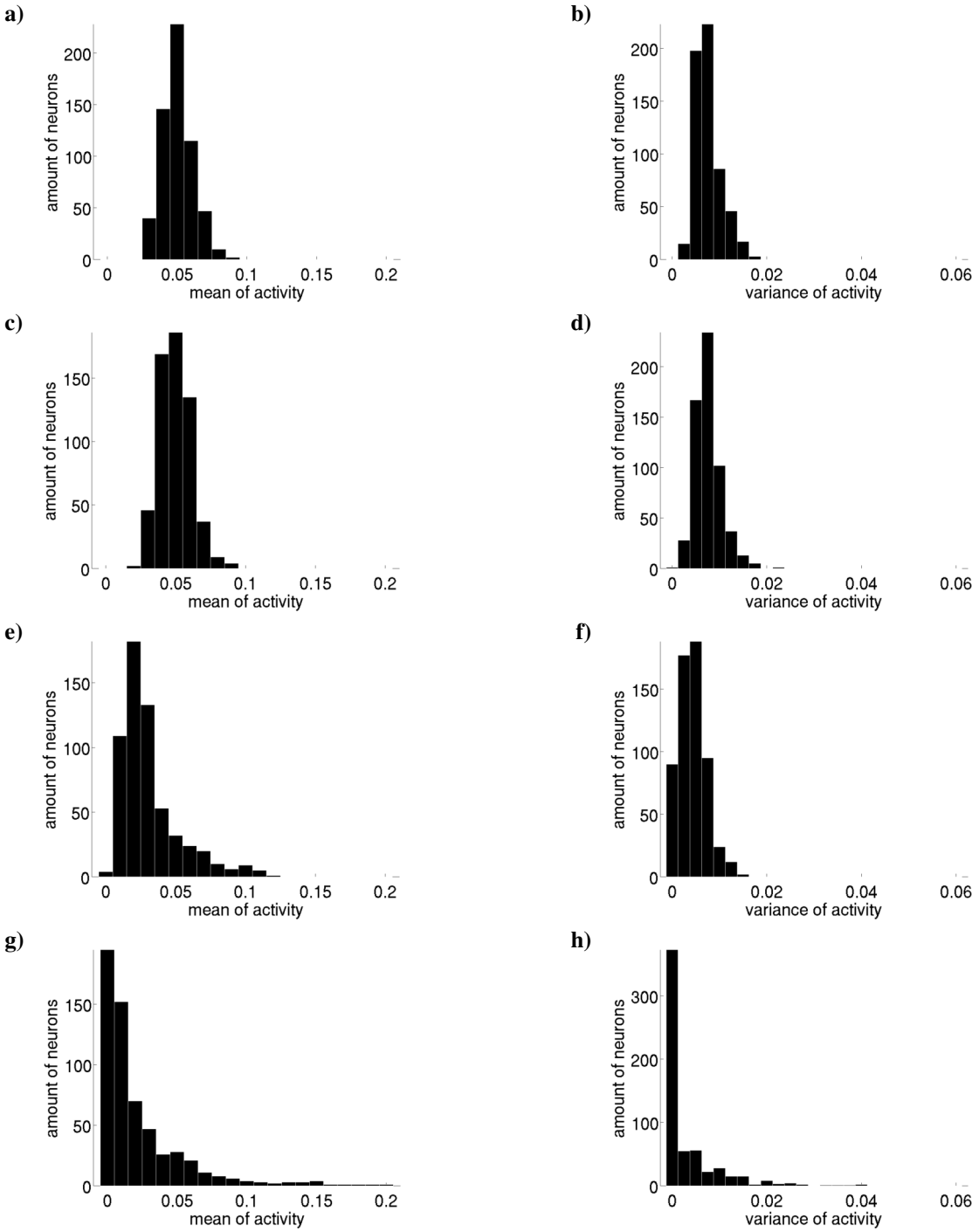

FIGURE 5.5: Histograms of the mean and variance of excitatory V1-layer 2/3 neurons activity, when regulating a single parameter in comparison to full and no intrinsic plasticity. The left column (a, c, e, g) shows the mean activity and the right column the variance (b, d, f, h). The first row (a, b) shows results for the model with full intrinsic plasticity, for comparison. In the second row (c, d) just the threshold θ is regulated. In the third row (e, f) just the slope a is regulated. The last row (g, h) shows the results for the model without intrinsic plasticity.

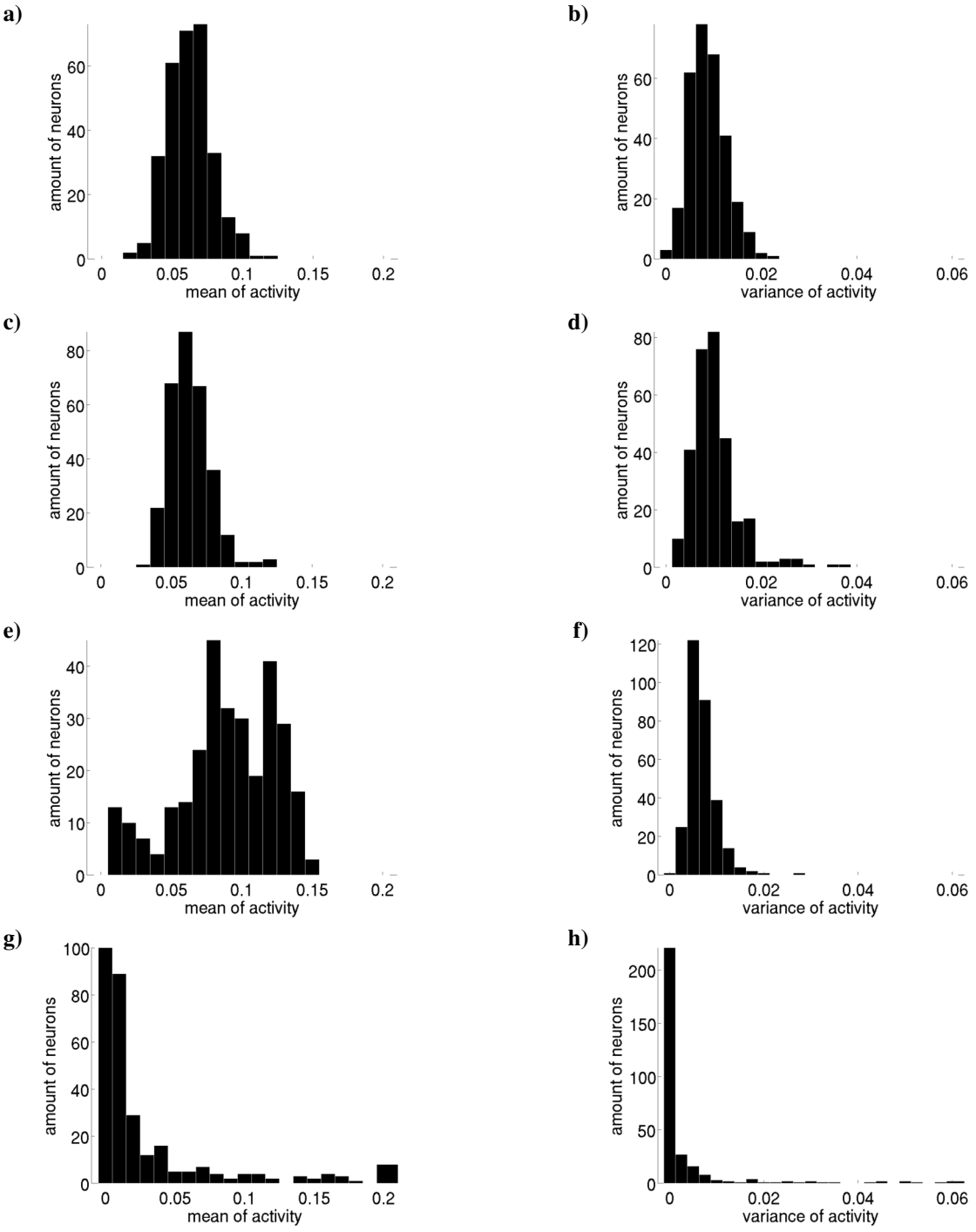

FIGURE 5.6: Histograms of the mean and variance of excitatory V2-layer 2/3 neurons activity, when regulating a single parameter in comparison to full and no intrinsic plasticity. The left column (a, c, e, g) shows the mean activity and the right column the variance (b, d, f, h). The first row (a, b) shows results for the model with full intrinsic plasticity, for comparison. In the second row (c, d) just the threshold θ is regulated. In the third row (e, f) just the slope a is regulated. The last row (g, h) shows the results for the model without intrinsic plasticity.

ulation of the mean has the largest effect on the distribution of the mean and the variance, probably caused by the inhibitory plasticity which is removing correlations between active neurons.

5.4.3 Encoding of visual objects by the neurons maximal information

To directly address the information encoding of the neurons, we calculated the information a neuron has about all classes (Rolls et al., 1997). Subsequently, we took the maximum of these values, which reveals if a neuron has useful encoding for anything in the dataset. Therefore, we calculated the information in bit each neuron has about each of the 100 object classes from the COIL-100 image dataset (for a dataset description see Sec. 6.3.5). We selected a central 24 by 24 patch from each of the 128 by 128 pixel large COIL-100 images and presented these patches to the network. We recorded the activities of all neurons onto each patch. Note, the network was not trained on any of these images. Further, all plasticity mechanisms have been turned off. We binned (0.05) these activities and calculated the following probabilities for the information measure (Eqn. 5.6).

$$I(\underbrace{s}_{\text{class}}, \overbrace{R}^{\text{activities}}) = \sum_r P(r|s) log_2 \frac{P(r|s)}{P(r)} \tag{5.6}$$

The probability that the neuron is firing with a certain activity is given by $P(r)$. $P(r|s)$ denotes the probability that the neuron fires with activity r when a stimulus from the class s is presented. $I(s,R)$ denotes the information the neuron conveys about class s given all responses R to each patch from this class. The maximal information a neuron can convey is given by the logarithm of the amount of classes (here $log_2(100) = 6.64$).

Each neuron has just a few high information values, when regarding the values for all classes. The most of the values are very low. Thus, the maximum value gives a good insight if a neuron encodes anything useful.

Suboptimal encodings can be revealed, when comparing the distribution of these maximal information values between models with and without intrinsic plasticity. For the excitatory neurons in V1-L4 we found similar distributions in both cases (Fig. 5.7ab). This has been expected from the largely similar mean and variance distributions of the neurons

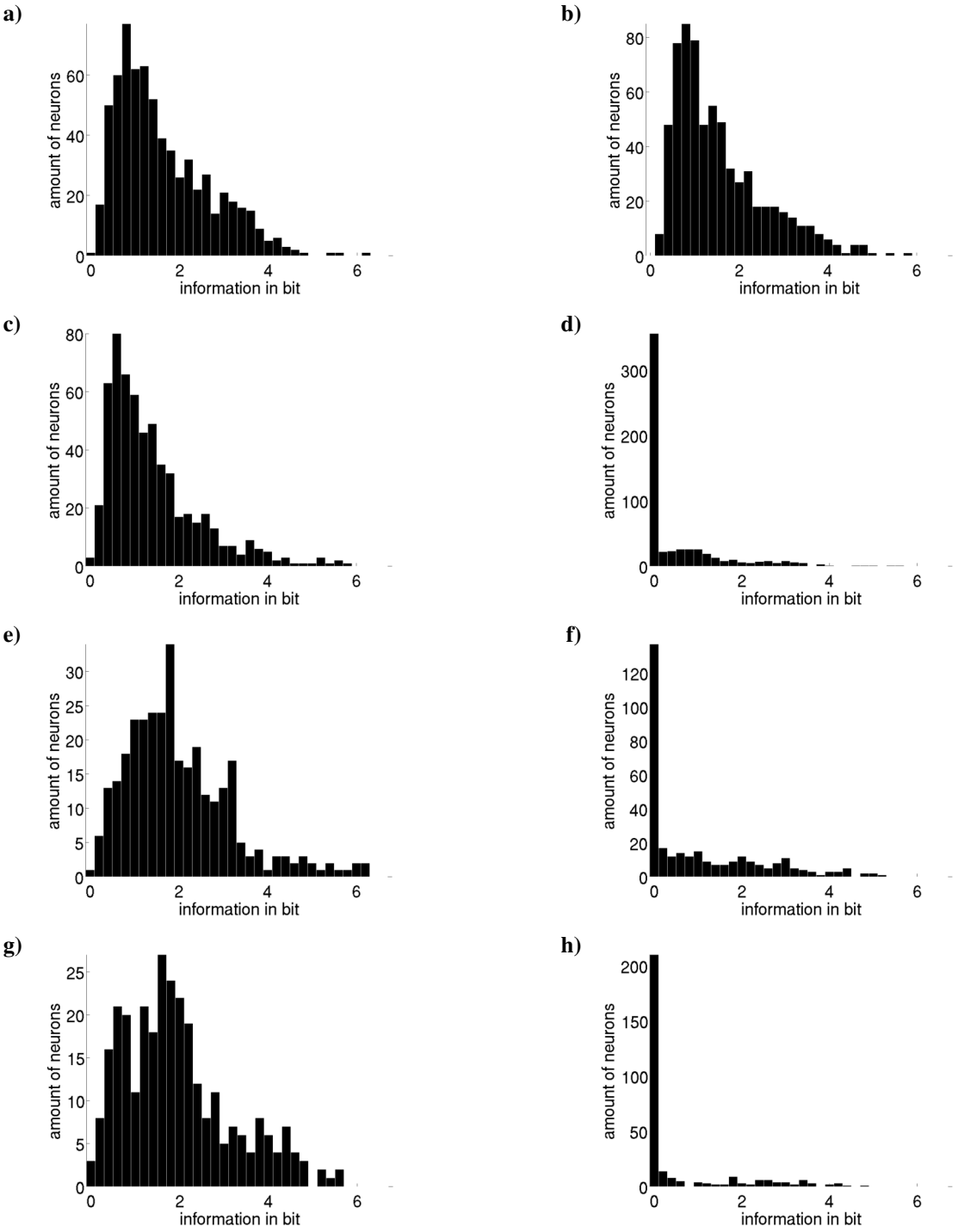

FIGURE 5.7: Maximal information of the neurons in the different layers, with and without inhibition. The left column shows results from a model with intrinsic plasticity, the right without intrinsic plasticity. The histogram of the maximal information values are shown for the excitatory neurons from early to deep layers. a,b) V1-L4; c,d) V1-L/23; e,f) V2-L4; g,h) V2-L2/3. While the in both models the first layer is similar, the most neurons in deeper layer of the model without intrinsic plasticity convey little to no information anymore.

(see Sec. 5.4.2). Further, the neurons in this layer have learned from the model inputs, which naturally have an equal activity distribution. Thus, large imbalances could not be expected. In the subsequent layer V1-L2/3 the information values of the excitatory neurons differ largely (Fig. 5.7cd). While the distribution with intrinsic plasticity is similar to the one of V1-L4, the distribution in the model without intrinsic plasticity is largely corrupted. The most neurons convey no or very little information. Once corrupted, the information values in the deeper layers of the model without intrinsic plasticity do not recover (Fig. 5.7fh). In contrast, the model with intrinsic plasticity shows a similar broad distribution as in the early layers, with slightly more neurons with higher maximal information values (Fig. 5.7eg). Thus, the neurons in deeper layers convey more information about a class than neurons in lower layers. This will be confirmed by the analysis of the recognition accuracies of the single layers (see Sec. 6.3.5).

5.4.4 Comparison of different drift strengths

The implemented intrinsic plasticity mechanism is a simple mechanism adapting the transfer function of a neuron to fulfill specific criteria of mean activity and variance. When these criteria are reached the intrinsic parameters become constant. However, the parameters are redundant to synaptic regulations, as synaptic scaling, which are also part of the model (see Sec. 6.2.3). That is, the length of the weight vector (synaptic regulation) can also change the mean activity and influences, together with the inhibition, the variance. Consequently, we integrated a preference for an unmodified activation function into the intrinsic regulations, a drift back to the origin of the regulation parameters (see Eqn. 5.5). When the intrinsic parameters become stable a small constant force is pulling them back to the unmodified values of threshold (zero) and slope (one). The drift force has to be chosen small enough to preserve the capability of regulating mean and variance. Further, the changes by the drift have to be slow enough that synaptic regulations can compensate the resulting response changes. On the other hand, the drift has to be fast enough to have a sufficient effect.

Hence, we compared different drift strengths and their influence on the parameter development within the network training. We can distinguish two phases of the network training, which indicate the sufficiency of the parameters. First, the phase of strong plasticity at the beginning of the network training, where a huge variability of the intrinsic parameters should be possible. Second, the phase after the network has converged (from

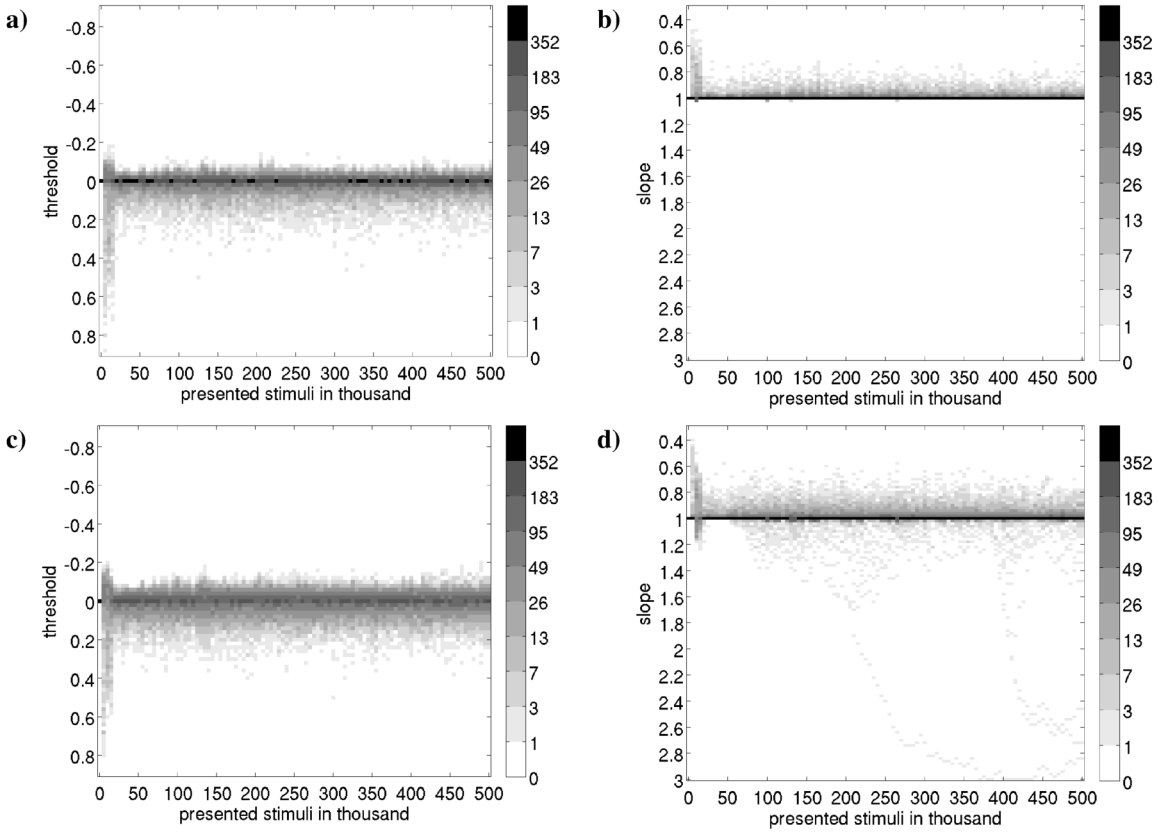

FIGURE 5.8: Development of the intrinsic parameters over training time with different drift speeds. The distributions of the intrinsic parameters threshold θ (left column) and slope parameter a (right column) are shown for the excitatory neurons in V1-L4 are shown, using different values of the drift parameter ε. From fast to slow these are a,b) $\frac{1}{50}$; c,d) $\frac{1}{100}$. The distribution is plotted every 5000 stimulus presentations during the network training. The gray value indicates the amount of neurons having a parameter value within the same bin (size 0.02). Note that we have used a logarithmic color map to enhance the visibility of the distribution. With increasing drift speed (bottom to top) the parameter value distributions become more scattered.

about the 100000th stimulus presentation), where the parameters should be declined to their origin. To gain insights in the model behavior we tested five different drift strengths ($\varepsilon = \frac{1}{50}$, $\frac{1}{100}$, $\frac{1}{250}$, $\frac{1}{500}$, and no drift; cf. Eqn. 5.5). We visualized the development of the distribution of the intrinsic parameters of the neurons in V1-layer 4 over the training time, other layers behave similar. Further, we calculated the fraction of neurons with a parameter value more than 0.1 apart from its origin over time.

In the default configuration of our model the drift speed is set to $\frac{1}{100}$ for the threshold as well as for the slope parameter. In the beginning of the training (the first 30000 stimulus presentations), a phase of strong changes in the weights, a large fraction of about

40 percent of the neurons developed threshold and slope parameters more than 0.1 apart from the origin (Fig. 5.8cd). Note that we have used a logarithmic color map to make the distribution around the origin visible. In the later training phase, the threshold values of the neurons cluster closely around the origin (less than 10 percent are outside the 0.1 range). Similarly, just about five percent of the slope values of the neurons had values more than 0.1 apart from the origin. When making the drift stronger (drift speed of $\frac{1}{50}$), we observed again many neurons with strongly changed parameters at the beginning of the training (Fig. 5.8ab). The threshold parameter was for 30 percent of the neurons more than 0.1 apart from the origin and the slope parameter for about 17 percent. In later phases of the training the most values are close to the origin. When using a drift speed of $\frac{1}{250}$, in the first training phase, about 35 percent of the threshold parameters and 75 percent of the slope parameters have been 0.1 apart of the origin. In the later training phase, about 10 percent of the threshold parameters and 25 percent of the slope parameters have been apart. With an even lower drift speed of $\frac{1}{500}$, we observed for the threshold parameter a similar amount (35 percent) of values more than 0.1 apart of the origin at the beginning of the training (Fig. 5.9a). However, in the later training phase more values than in the default configuration, about 15 percent, lay apart from the origin. The fluctuation of the slope parameter became much stronger (Fig. 5.9b). In the beginning about 90 percent have been apart from the origin and in the later training phase about 50 percent have been apart. When completely deactivating the drift about 75 percent of the threshold values have been apart of the origin in the beginning of the training (Fig. 5.9c). In the later training phase about 15 percent have been apart. For the slope parameter. 90 percent have been more than 0.1 apart from the origin in the beginning and later a high value of about 70 percent remain apart (Fig. 5.9d). When using no or very slow drift speeds we observed in a few deep layers (e.g. the inhibitory neurons of V2-L2/3) distributions of slope parameters largely apart from the origin. When using faster drift speeds (e.g. $\frac{1}{250}$) this has not been observed.

When no drift mechanism was used the slope values end widely distributed between values of 0.3 and 2, whereas the threshold values cluster in 0.2 range around the origin. Hence, a drift seems not mandatory to regulate the threshold value. Nevertheless, the chosen default drift value ($\frac{1}{100}$) pushes the majority of threshold parameters close to the origin and allows a sufficient dynamic range in the first phase of training. Contrary to the threshold, the drift seems mandatory for the slope parameter. Neurons in deep layers can develop exceedingly strong parameter values and the majority of the parameter values do

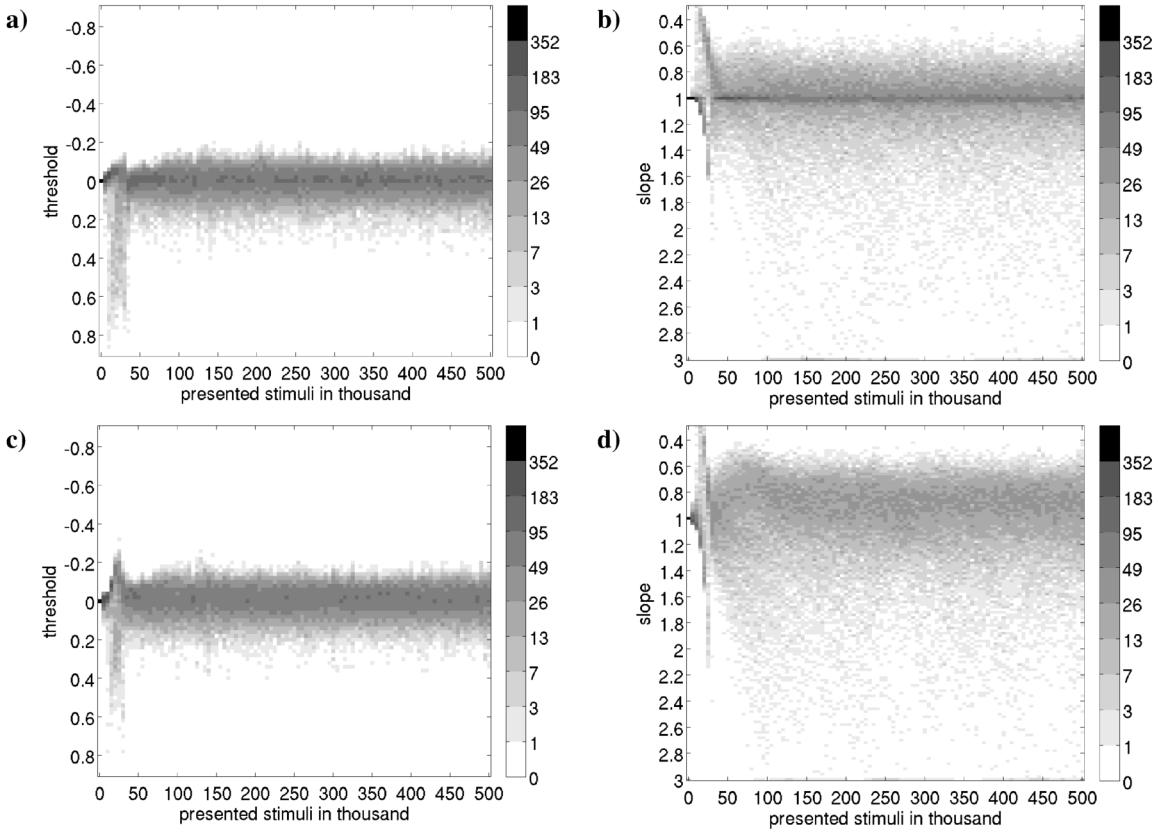

FIGURE 5.9: Development of the intrinsic parameters over training time with different drift speeds. The distributions of the intrinsic parameters threshold θ (left column) and slope parameter a (right column) are shown for the excitatory neurons in V1-L4 are shown, using different values of the drift parameter ε. From fast to slow these are a,b) $\frac{1}{500}$; c,d) no drift. The distribution is plotted every 5000 stimulus presentations during the network training. The gray value indicates the amount of neurons having a parameter value within the same bin (size 0.02). Note that we have used a logarithmic color map to enhance the visibility of the distribution. With increasing drift speed (bottom to top) the parameter value distributions become more scattered.

not decline to values close to the origin. With drift speeds of $\frac{1}{250}$ and slower many neurons end with slope parameter not fulfilling the goal of a minimal modified activation function. When using a too fast drift speed as $\frac{1}{50}$ the regulation of the slope parameter was impaired. Thus, the default parametrization with $\frac{1}{100}$ appears suitable.

5.4.5 Activity distribution

An exponential distribution of activities is suggested to be optimal, for a fixed mean, in terms of information transmission (Simoncelli and Olshausen, 2001) and energy efficiency

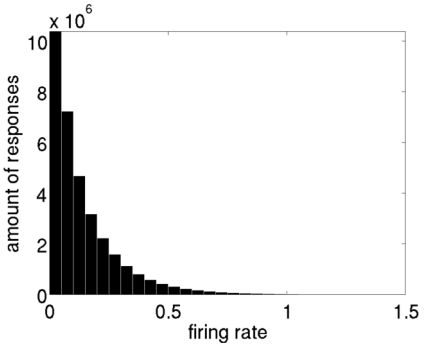

FIGURE 5.10: Distribution of neuronal activities of V1-layer 4. The histogram shows the non zero activities of all neurons in V1-layer 4 on 100000 natural scene stimuli. The activities are exponentially like distributed, without having a constraint enforcing such a distribution.

(Rolls and Treves, 2011; Levy and Baxter, 1996). This is argued for other implementations of intrinsic plasticity (e.g. Stemmler and Koch, 1999; Triesch, 2005a,b, 2007; Savin et al., 2010). With that in mind Triesch (2005a) introduced a method which shapes the activation function of each neuron to obtain exponentially distributed responses. As activation function he used an configurable sigmoidal function. The parameters for this function are found during learning by a stochastic gradient descent approach minimizing the Kullback-Leibler divergence of the neuronal responses to an exponential function. Thus, an exponential distribution of the activities is explicitly targeted by this implementation. Earlier, Stemmler and Koch (1999) introduced a Hodgkin-Huxley model with voltage dependent conductances. The conductances have been adapted to achieve a particular average firing rate. This changes the activation function in a way that it covert a Gaussian input distribution into an exponential response distribution. Similarly Triesch (2005b) suggested the adaption of a sigmoidal activation function to match the mean and variance of a given exponential distribution. These approaches implement their mechanisms on single cell level, however, an exponential response distribution alone can not grant optimality of information transmission in a multi neuron setup (Simoncelli and Olshausen, 2001).

In contrast to Triesch (2005a) our intrinsic plasticity implementation has no objective constraining the responses to an exponential distribution. Moreover, also our other plasticity mechanisms have not such an objective (see Chapter 6). Indeed, inhibition could shape the response distribution in that way. However, its anti-Hebbian learning mechanism let the synaptic weights develop relative to the neurons correlation (see Sec. 6.2.4), so that inhibitory learning implements also no obvious mechanism to constrain the response distribution, beside reducing correlations between the neurons. We speculated that the observed exponential distribution of the neuronal responses (Baddeley et al., 1997) is just a byproduct of the natural interplay between an almost linear activation function (Ringach

and Malone, 2007) and Hebbian excitatory and inhibitory learning. When this is the case our network should also show exponentially distributed responses.

Hence, we measured the response distribution of our model neurons to natural scene stimuli. Therefore, we recorded the responses of all model neurons to 100000 natural scene stimuli and measured the ratio between standard derivation and mean response, also known as the coefficient of variation (CV). For an exponential distribution this simple measurement should give values close to one. We excluded zero responses from the evaluation as they are no valid values for an exponential function. Further, we visualized the distributions for each neuron population. We controlled visually the exponential character of the distributions by plotting the data on a semilogarithmic scale where the occurrence of a value should change linearly (Triesch, 2005b, 2007; Stemmler and Koch, 1999). Finally, we compared our results to the data of Butko and Triesch (2007), which explicitly aim exponentially distributed activities in a V1 simple-cell learning task. Similar to other publications of their lab, this done by minimizing the Kullback-Leibler divergence of the neuronal responses to an exponential function via parameter adaption of the used sigmoidal activation function. For comparison, we calculated the similarity between the response data of our V1-layer 4 neurons to a generated exponential distribution having the same mean as our data. Therefore, we binned our data and the desired exponential distribution in 50 equally spaced bins and calculated the L2-norm between both distributions analogue to Butko and Triesch (2007). Whereas Butko and Triesch (2007) reported the development of the similarity, we compare just the similarity after training.

We found that the responses of all neuron populations follow an exponential distribution. The coefficients of variation have been close to one (Table 5.1). This result has been visually confirmed by the shape of the distribution (for V1-layer 4 see Fig. 5.10). Further, we found a linear slope when plotting the data on a semilogarithmic scale (results not shown). The similarity to an exponential distribution (L2-norm) for our excitatory V1-layer 4 neurons was 0.02, which is in the range between the model versions of Butko and Triesch (2007, Fig. 7) with weak intrinsic plasticity, “Low IP”, and strong intrinsic plasticity, “High IP”. It is much closer to an ideal exponential distribution as their case without intrinsic plasticity ($\sim$0.2), “No IP”, and their case “Linear” ($\sim$1.0), which uses a linear activation function instead of a sigmoidal one and, of course, has no intrinsic plasticity. To underpin our opinion that intrinsic plasticity is not needed to achieve exponentially distributed activities, we measured the similarity also for our model without intrinsic plas-

Layer	Type	CV
V1-L4	Excitatory	1.0658
	Inhibitory	1.1541
V1-L2/3	Excitatory	0.9842
	Inhibitory	1.0351
V2-L4	Excitatory	0.9409
	Inhibitory	1.0528
V2-L2/3	Excitatory	0.8972
	Inhibitory	1.0154

TABLE 5.1: Coefficient of variation for all learned neuron populations. The table gives an overview of the coefficient of variation (CV) for the responses on 100000 natural scene stimuli of all neuron populations having plastic synapses. A value close to one indicates an exponential distribution. The CV value for all populations indicate that they have an exponential like distribution of their responses. The values are averages over 10 model runs. The standard derivation was 1 percent and below for all populations, except V2-layer 2/3, having the least amount of neurons, had a standard deviation below two (excitatory) and five (inhibitory) percent.

ticity. Despite the poor learning results in deeper layers, the first layer (V1-layer 4) should give us valid results. The excitatory neurons in that layer achieved a similarity of 0.04, again better than the "No IP" case of Butko and Triesch (2007, Fig. 7) and comparable to the "Low IP" case. Note, that the exact values of the similarity measure are hardly comparable as the parameters of the discretization influences the results. However, we achieved qualitatively similar results when using different binnings.

The exponential character of our model responses, without having such an objective, encourages our assumption that exponential activity distributions can not be rendered as objective of single brain parts, as the neurons. They are rather the byproduct of the recurrent brain circuit, we modeled here, and need no explicit objective. Indeed, the use of our intrinsic plasticity mechanism improves the similarity to an exponential distribution in comparison to the model version without intrinsic plasticity. But the neuronal responses in both models showed a high level of similarity. Thus, our intrinsic plasticity mechanism can not be counted to implicitly implement an objective for exponential activity distributions. The more likely source for the exponential character of the responses is the inhibitory plasticity. Our inhibitory plasticity decorrelates the neuronal responses, which in turn sparsifies the code and might cause an improve in similarity to an exponential distribution. More important, decorrelation enforces independence of the neuronal responses which is the inevitable criteria for efficient information encoding in a system of multiple neurons

(Simoncelli and Olshausen, 2001), in contrast to exponentially distributed responses (for a fixed mean) in the single neuron case. Similarly to our intrinsic plasticity mechanism, Stemmler and Koch (1999) and Triesch (2005b) showed that adapting the mean firing rate in a single neuron setup can shape a Gaussian input distribution to an exponential response distribution. However, we use a rectified linear function which is more limited in its capability to change the distribution and we showed that also without using intrinsic plasticity a high degree of similarity can be obtained.

5.5 Conclusion

We have shown that intrinsic plasticity can balance the mean and variance of the neurons. The distributions of these activity moments became narrower, which largely improved the information coding in deeper layers. An improvement was also achieved when either threshold or slope of the activation function was regulated. However, the regulation of just the threshold appeared superior in comparison to regulating just the slope. The means have been narrower distributed while the distribution of the variances are sufficient narrow.

Further, we implemented a preference for a minimal modified activation function by a constant drift to the regulation parameters origin. We showed that with this drift a minimal modified activation function can be achieved. We compared different strengths of the drift speed to gain insights whether the dynamic of the intrinsic regulation is impaired and the goal of a minimal modified activation function is achieved. We found that a drift speed of $\varepsilon = \frac{1}{100}$ allows a sufficient parameter dynamic while the majority of neurons developed a minimal modified activation function at the end of the network training.

Finally, we found that our neuron responses are exponentially distributed, despite we have no objective for it. In contrast to previous approaches, we do not attribute this to the intrinsic plasticity mechanism. This was underpinned by the finding of an exponential distribution of the network responses in a model without intrinsic plasticity. We attribute the response distribution in our model to the inhibitory plasticity, which improves information encoding by enforcing independence of the neuronal responses.

6 Synaptic Plasticity and Homeostatic Regulations

In this chapter we introduce the last part of our model, the synaptic plasticity. First, we introduce the fundamental principles of synaptic plasticity, followed by relevant computational principles and models. Then, we present our implementation. Subsequently, we evaluate the response properties and learned connectivity of the neurons. Finally, we show the capability of our network for invariant object recognition.

6.1 Introduction

Synaptic plasticity is the core principle for a self-organizing model of the visual system. However, several different plasticity rules have been proposed in the last decades. The methods largely differ in their degree of biological detail and theoretical concepts. Following, we will introduce the fundamental principles to which our mechanisms are related. Subsequently, we describe the synaptic plasticity model which we have derived from these principles (Sec. 6.2). We evaluate the resulting model, including all previously introduced plasticity mechanisms as before, for different aspects (Sec. 6.3). First, we show that our neuron responses follow typical criteria of efficient coding. Than, we investigate the learned receptive fields of the neurons and compare the V2 responses to a novel hypothesis for its feature sensitivity. Further, we analyze the learned weight distributions and the relation between weights and response correlations. We demonstrate the relation between synaptic plasticity and structural plasticity, in terms of analyzing the connection probabilities in relation to the correlation between the neurons. Finally, we show the learned invariance properties and the ability of the model for object recognition.

6.1.1 Hebbian and anti-Hebbian learning

Mathematical models of synaptic plasticity can be traced back to the fundamental formulation of Donald Hebb, which points out that the coincidence of the firing of two neurons causes the strengthening of their connection (Hebb, 1949). Neuroscientific research found later two primary mechanisms changing the synaptic strength: long term potentiation (LTP), increasing the strength, and long term depression (LTD), decreasing the strength (Malenka and Bear, 2004; Feldman, 2009). Other research revealed the functioning of the first layers in the visual cortex and described their receptive fields (Hubel and Wiesel, 1962; Jones and Palmer, 1987), which emerge through visual experience (Ohzawa and Freeman, 1988) and synaptic plasticity mechanisms.

In the primary visual cortex (V1) layer-4 the so called simple-cells have been found. These cells could be easily described by maps of excitatory and inhibitory effects of stimulation with light (Hubel and Wiesel, 1962). Many computational models could show that the found receptive fields can be reproduced by applying some core principles on natural scene images as sparseness (Olshausen and Field, 1996, 1997; Rehn and Sommer, 2007; Rozell et al., 2008; Weber and Triesch, 2008), finding independent components (Bell and Sejnowski, 1997; van Hateren and van der Schaaf, 1998; van Hateren and Ruderman, 1998), or Hebbian learning in combination with anti-Hebbian decorrelation (Falconbridge et al., 2006; Wiltschut and Hamker, 2009; Zylberberg et al., 2011; King et al., 2013). With that, the demonstration of the algorithm capabilities on learning V1 simple-cell like receptive fields became a testbed.

As described in the General Introduction, we focus on the methods using synaptic plasticity in terms of rate based Hebbian learning. From the postulate of Hebb (Hebb, 1949) a naive learning rule can be derived. This is the product of the pre- (r_i) and postsynaptic activity (r_j). However, this explains just long term potentiation (LTP). The weight w_{ij} can only increase for positively defined firing rates.

$$\frac{dw_{ij}}{dt} = r_i r_j$$

A solution to account for long term depression (LTD) is the introduction of a pre- or postsynaptic threshold θ on the activity (Dayan and Abbott, 2001, Sec. 8.2). When the threshold is taken close to the average activity of the related neuron the weight change becomes relative to the covariance of the activities (Dayan and Abbott, 2001; Sejnowski,

1977), and with that also relative to the response correlation between the neurons. Unfortunately, this formulation also suffers from an unbound increase of the synaptic weights as the naive Hebbian rule. That is, when the presynaptic activity positively correlates with the postsynaptic activity, than the weight will increase, which in turn increases the response correlation.

$$\frac{dw_{ij}}{dt} = r_i(r_j - \theta_j)$$

A potential solution was originally developed for the naive Hebbian term, but can also be applied to the covariance learning rule. Oja (1982) proposed a term just dependent on the postsynaptic activity and the regarded synaptic weight, which normalizes the weight relative to all weights. This is possible because the postsynaptic activity results from the whole weight vector, hence, its height indicates whether the weights should be overall increased or reduced to achieve the desired activity level. The weight change from this term is relative to the height of the regarded weight, which preserves the relation between the neurons. Therefore, this normalization is called a multiplicative normalization.

$$\frac{dw_{ij}}{dt} = r_i r_j - r_j^2 w$$

The combination of this multiplicative normalization, we will call "Oja normalization", with covariance learning (Wiltschut and Hamker, 2009) or the naive Hebbian term (Falconbridge et al., 2006) has been successfully applied for learning V1 simple-cell receptive fields.

Another well established synaptic plasticity rule is the so called BCM rule (Bienenstock et al., 1982). This rule shares high similarities with a covariance rule, having the threshold on the postsynaptic side. Additionally to a covariance rule, the BCM rule multiplies the weight changes with the postsynaptic activity. This results in weak changes for low postsynaptic activities and strong changes for high activities. In comparison to covariance and naive Hebbian rule, this rule uses a dynamic threshold to prevent an unbound increase of the synaptic weights. A typical method to calculate the threshold uses the temporal average of the squared postsynaptic activity (Law and Cooper, 1994; Intrator and Cooper, 1992). This type of normalization increases the amount of LTD when the postsynaptic activity increases. That is, when the threshold increases with activity than the zones of weight decrease is extended and the zone of increase is confined. As consequence, the neurons can just increase their weights for a subset of stimuli, which give in total the a higher amount

of weight increase as all stimuli, which evoke weak activities, give as weights decrease. We took the BCM equation in the version of Willmore et al. (2012).

$$\frac{dw_{ij}}{dt} = r_i(r_j - \theta_j)r_j$$

The BCM rule has also been used to learn V1 simple-cell receptive fields. In Law and Cooper (1994) the formation of single receptive fields has been demonstrated. Willmore et al. (2012) combined the rule with a normalization term on the postsynaptic activity, which puts the activity in relation to the activities of all other neurons. This implements a kind of inhibition which lets the neurons just increase their weights when they are more active than other neurons. Because of that, a population of neurons which covers the input space with different receptive fields developed.

To work in a network and encode different properties of the input also the naive Hebbian rule and the covariance rule need an inhibitory mechanisms, which induces competition between the neurons for the patterns to they represent. A promising strategy has been proposed with local anti-Hebbian learning (Földiák, 1990). A network of several neurons, learning with a Hebbian learning rule, was extended with lateral connections. The weights of these lateral connections c_{kj} have been learned with the anti-Hebbian rule. This increases the inhibitory weight when both neurons are active, similar to the naive Hebbian term. The weight decreases by the amount the product of the activities would give when both units are uncorrelated, this is the product of the mean activities $(\bar{r}_k, \bar{r}_j)$. In consequence of this formulation, the inhibition between the neurons increases until they are uncorrelated, thus, the inhibition enforces decorrelation. We show the formulation of King et al. (2013), Földiák (1990) used directly the knowledge about the probability of the activities.

$$\frac{dc_{kj}}{dt} = r_k r_j - \bar{r}_k \bar{r}_j$$

Földiák (1990) demonstrated that the neurons in the network can learn the independent components of bars presented at random positions, which are again the bars at the different positions. Further, he point out that all components of an image can be represented in parallel. This means, when presenting more than one bar then all units responding on one of these bars will be active in parallel.

Wiltschut and Hamker (2009) adapted this concept and made the learned weight relative to the correlation of the neurons. The weight becomes low for uncorrelated neurons and

high for correlated one. The parameter α controls how high the weight can grow and can be interpreted as a multiple of the average activity of presynaptic neuron. This rule reduces weights, in contrast to Földiák (1990), between neurons which are not correlated anymore, but have been correlated in the course of the learning.

$$\frac{dc_{kj}}{dt} = r_k r_j - \alpha r_j c_{kj}$$

An extension to this rule has been independently developed by King et al. (2013). To achieve a correct alignment with the covariance between the neurons they corrected the weight by a baseline and used the average activities directly. The advantage to the rule of Wiltschut and Hamker (2009) is that the weight value becomes zero when the linear correlation (Pearson) of both neuron activities is zero.

$$\frac{dc_{kj}}{dt} = r_k r_j - \bar{r}_k \bar{r}_j (1 + c_{kj})$$

Vogels et al. (2011) introduced a STDP rule for inhibitory synapses which turns out very similar to Földiák (1990). The spike event dependent rule can be summarized to following description, which is also applicable in rate based models.

$$\frac{dc_{kj}}{dt} = r_k r_j - r_k \rho_0$$

The parameter ρ_0 determines the target activity for the postsynaptic neuron. That is, the amount of inhibition the neuron receives is regulated by the reduction term in the learning rule. Instead of the average activity of the presynaptic neuron the presynaptic activity is used directly. This has the similar effect, as in the rule of Wiltschut and Hamker (2009), the slow speed of the weight change implements a temporal trace over the presynaptic activity in this term, which approximates the average activity. Also when this rule appears like a reinvention of Földiák (1990) it has the advantageous property of controlling the postsynaptic activity and with that it implements a form of firing rate homeostasis. Indeed, it shares the same drawback of remaining high weight values for decorrelated neurons.

6.1.2 Trace learning

V1 simple-cells became the standard testbed for models of synaptic plasticity. However, in V1 also the so called complex-cells have been found. Their most remarkable property is the similar response characteristic to simple-cells, while being slightly independent to the precise stimulus position (Hubel and Wiesel, 1962). With that property, these cells became the role model for recognition invariance in the cortex. When we speak about invariance in the visual system, we mean an increased tolerance, in terms of similar neuron responses, to changes in the appearance of objects or the features where the objects are composed from. Typical changes in the appearance are changes in scale, rotation, and position in the scene, but also all changes depending on the viewpoint, illumination, or the texture. However, just a few synaptic plasticity rules are proposed for the learning of complex-cell like invariance properties.

One main idea for learning invariance properties is that the world, the visual scene, changes temporally slow, on the contrary, the retinal image can vary rapidly (Berkes and Wiskott, 2005; Teichmann et al., 2012). The visual cortex has to somehow compensate this rapid variations, thus, it has to enforce a temporally slow changing neuronal code. Several computational approaches used an objective function dependent on temporal slowness to minimize the differences of the outputs to obtain complex-cell like response properties (Kayser et al., 2001; Hashimoto, 2003; Körding et al., 2004; Berkes and Wiskott, 2005). Another approach is to modify Hebbian synaptic plasticity by replacing the postsynaptic activation with a history of the neuron activations (Földiák, 1991; Wallis and Rolls, 1997; Einhäuser et al., 2002; Spratling, 2005; Teichmann et al., 2012). These approaches require temporal coherent input, i.e. consecutive images have to be strongly related to each other. Which should be, of course, the case in the real world. An alternative approach to the use of the history of activities was introduced with continuous transformation learning (Stringer et al., 2006; Stringer and Rolls, 2008; Perry et al., 2010; Evans and Stringer, 2012; Rolls, 2012). It exploits the spatial coherence of the input images, instead of the temporal. This means, it assumes that when the changes between the images are low, many neurons should be active for different views of the same object. This sustained activity is proposed to have the same effect as using an activity history, because active neurons can learn to connect to all sets of neurons representing an object. While we here will focus on rules which are proposed as Hebbian synaptic plasticity rules, we have to mention that there are also algorithms which do not exploit any form of temporal coherence, but are also not proposed

as rules for synaptic plasticity (e.g. Hyvärinen and Oja, 2000; Hyvärinen and Hoyer, 2001; Osindero et al., 2006; Karklin and Lewicki, 2009; Köster and Hyvärinen, 2010). For a broader introduction into the different approaches see our previous work Teichmann et al. (2012).

Since we aim to implement a form of synaptic plasticity, we will focus on the algorithm class using an activity history to exploit the temporal coherences in the input images for learning. This activity history can have a very simple form as using the activity on the previous image while learning from the current inputs (Spratling, 2005). More common is the usage of a trace over the past activations. This trace stores the previous activities with a certain decay and adds the current activity (Földiák, 1991). The physiological counterpart to the trace might be the postsynaptic calcium concentration. In our previous work Teichmann et al. (2012), we introduced the role of calcium for synaptic plasticity as following. "It has been shown that many forms of bidirectional synaptic plasticity (long term potentiation (LTP) and long term depression (LTD)) are calcium dependent (Malinow et al., 1989; Daw et al., 1993; Pettit et al., 1994; Lledo et al., 1995; Hu et al., 2001). Calcium/calmodulin-dependent protein kinase (CaMKK) signaling following NMDA receptor activation has been identified as a cell-autonomous homeostatic regulator of synaptic strength in response to activity (Goold and Nicoll, 2010) such that potentiation and depression of synaptic connection strength depends on the level of calcium at the corresponding synapses (Cummings et al., 1996; Yang et al., 1999; Cho et al., 2001; Cormier et al., 2001). These and other electrophysiological studies have lead to a framework of calcium-based learning (Lisman, 1989; Shouval et al., 2002b,a) which suggests that the intracellular calcium concentration influences the strength and temporal dynamics of neuronal learning (Shouval et al., 2002b)."

From these ideas we developed in Teichmann et al. (2012) a calcium dependent Hebbian learning rule. We implemented a temporal trace following the neurons activity, which represents the calcium concentration or calcium level. We replaced the activities in the Hebbian learning rule of the excitatory synapses by the calcium levels. The learning rule further combined the principle of covariance learning with Oja normalization. Furthermore, the time constant for the weight change, i.e. learning speed, was also calcium dependent. That is, for low postsynaptic calcium levels we learned slower than for high levels. We demonstrated the capability of this learning rule in a model learning V1 complex-cell receptive fields. We used Gabor filters of different orientations and positions to resemble V1

simple-cells. As input we used natural scenes. We showed that with this approach the majority of neurons developed complex-cell like receptive fields. The neurons showed shift invariance and the neuron responses are found to change slower for varying inputs. Additionally, we tested whether this complex-cell properties could be achieved without trace learning. Therefore, we used a very fast trace, short in comparison to the presentation time of a single stimulus. With this fast trace we could classify just the half amount of neurons as complex-cells in comparison to the long trace. We further controlled the influence of the temporal coherent input. We found that with a random presentation protocol the amount of complex-cells dropped from about 90 percent to the half, similar to the effect of the fast trace, for with the temporal order played no role.

Despite the higher complexity of our trace learning rule, it shares large similarity to the principle proposed by Földiák (1991). He also used a trace following the postsynaptic activity in a multiplicative way. This trace replaced the postsynaptic activity in a Hebbian learning rule. However, while we used lateral inhibition for the decorrelation of the units, he used a winner-take-all mechanism. The network was trained with moving lines of four different orientations. The neurons in the network became selective to lines of a single orientations, regardless of their position. He also tested the learning without using a trace and found that the neurons tended to represent more than a single orientation and had been more sensitive to the positions of the lines.

Wallis and Rolls (1997) again used the rule of Földiák (1991) in a far deeper network with four layers. They demonstrated the capability of this learning principle on more complex stimuli, such as faces. Further, they showed that invariance gradually increased over the hierarchy.

6.2 Synaptic plasticity and homeostatic regulations methods

6.2.1 Neuronal calcium level

Based on our previous work (Teichmann et al., 2012) the learning in our network is calcium dependent. This means that instead of the postsynaptic activity we are using the postsynaptic calcium concentration Ca_j in the learning rule. This calcium concentration

follows the firing rate r_j of the neuron j (Eqn. 6.1) and implements a trace over its activity.

$$\tau_{Ca}\frac{dCa_j}{dt} = r_j - Ca_j \tag{6.1}$$

We distinguished between two speeds of the trace. One fast ($\tau_{Ca} = 10ms$) for excitatory neurons in the layer-4 of an area and all inhibitory neurons. This should allow feature extraction based on the statistic of the present inputs. The fast trace is chosen that fast that no differences in the learned receptive fields are observed in comparison to a model using directly the firing rate. The slow trace ($\tau_{Ca} = 500ms$), applied for the excitatory neurons in layer-2/3, is much slower than the presentation time of a single stimulus ($100ms$), so that it causes that neurons learn from subsequent input stimuli, instead only the present one as in Hebbian learning. This, together with simulated eye movements for inducing temporal dependencies in the input (see Sec. 3), enables the model to learn invariant representations. The difference in learning between the layers follows the established assumption about the processing in the ventral stream that feature extraction is followed by a step building invariance in each area (Riesenhuber and Poggio, 1999; Serre et al., 2007).

6.2.2 Time constant for calcium dependent synaptic change

The neuronal calcium level Ca_k also influences the speed of the synaptic weight change $\tau_{Learn,j}$. In biological recordings a higher calcium level caused a higher alteration of the synaptic efficiency (Shouval et al., 2002b). We implemented this, similar to Teichmann et al. (2012), by a function leading to high values (slow) of $\tau_{Learn,j}$ for low calcium values, which exponentially decays with increasing calcium level (Eqn. 6.2).

$$\tau_{Learn,j} = a + b \cdot e^{\left(-c \cdot Ca_j^{post}\right)} \tag{6.2}$$

The parameters a and b define the lower and upper bound of the learning speed and c defines the exponential decay (for values see Table A.5).

6.2.3 Calcium dependent Hebbian learning of excitatory connections

Calcium dependent synaptic change

We modified our previously proposed learning rule (Teichmann et al., 2012) to achieve more simplicity. In Teichmann et al. (2012) we used a learning rule based on the one of Wiltschut and Hamker (2009), which is combining the ideas of covariance learning (Dayan and Abbott, 2001; Sejnowski, 1977) and normalization Oja (1982). We extended this rule by a calcium trace and calcium dependent learning speeds (Shouval et al., 2002b; Yeung et al., 2004; Castellani et al., 2005). Further, we applied the used Oja normalization just when having a sufficiently high amount of postsynaptic calcium concentration. The Oja normalization constrains the length of the weight vector (L2-norm) relative to the value of the parameter α. With that learning rule we have been able to learn V1-complex like neuron properties as shift and phase invariance while being orientation selective (Teichmann et al., 2012). To be able to learn this invariance properties, we calculated a calcium trace over the postsynaptic activity of the neuron (see above). This calcium trace exploits the circumstance that the visual scenery, i.e. the entities composing the image, underly slower changes than the retinal image. Thus, with trace learning neurons learn from subsequent inputs representing the same scenery which results in an invariance against rapid input changes.

Here, we simplified our new learning rule in comparison to the previous one by removing the covariance term on the postsynaptic term, so that it now relies on the postsynaptic calcium concentration Ca_j only. This removes the case separation between high and low postsynaptic calcium concentrations and we can always apply the normalization term. The normalization term in turn bases on the postsynaptic calcium concentration, we removed the covariance term accordingly. The α value, defining the length of the weight vector, is, as previously, regulated dynamically (see next Section). On the presynaptic term we now use the neuronal activity r_i, instead of the calcium concentration, together with a covariance term. That is, the presynaptic activity is reduced by a threshold to directly account for long term depression (LTD), which has been induced in experimental setups by low frequency stimulation (Malenka and Bear, 2004). Using the neurons average presynaptic activity appears to be a good choice for the threshold value, in a sparse code it is sufficiently low and should give a good indicator to distinguish between pattern of interest

and noise. However, the parameter has to be determined in advance and has to operate at the beginning of network training, where the network activity largely differs from later phases, as well as in the later phases. For simplicity we chose the average activity $\bar{r}^{pre}$ of the presynaptic population (similar to Wiltschut and Hamker, 2009; Teichmann et al., 2012), which is the same as we chose for the target activity in the intrinsic plasticity rule (see Sec. 5.3). The new learning rule, calculating the change of the synaptic weight w_{ij} for the excitatory synapses of our model, reads as follows (Eqn. 6.3). Note, as all other connections the weights are defined positive.

$$\tau_{Learn,j}\frac{dw_{ij}}{dt} = (r_i - \bar{r}^{pre}) \cdot Ca_j - \alpha_j \left(Ca_j\right)^2 w_{ij} \tag{6.3}$$

with

$$w_{ij} = (w_{ij})^+$$

Without postsynaptic threshold the differentiation of the neurons occurs over their different activities, i.e. the postsynaptic calcium concentration. This means that neurons have strong weight changes for patterns where they strongly respond, while other neurons strongly respond for other patterns and learn these different patterns. Therefore, a differentiation of the activities is needed. This is facilitated by the inhibitory plasticity (see Sec. 6.2.4), which increases inhibition between similarly responding neurons and as a consequence increases the differences of their activities. This might mean that just one of two competing neurons remains active. We stabilized this through intrinsic plasticity, which reinforces neurons with weaker activities and down regulates neurons with strong activities (cf. Sec. 5.3).

Homeostatic regulation

Two different kinds of homeostatic mechanisms have been found, a multiplicative change of synaptic efficiency and a change in intrinsic excitability and activation threshold. The later we address with our intrinsic plasticity mechanism (Chapter 5.3). The first one is part of the here introduced synaptic plasticity model.

The Oja normalization (see above) multiplicatively restricts the length of the weight vector to $\frac{1}{\alpha}$ (Dayan and Abbott, 2001, Sec. 8.2). It does not account for the observed changes induced by long periods of low, or high, presynaptic activity (Turrigiano et al., 1998; Desai et al., 1999; Turrigiano and Nelson, 2004; Nelson and Turrigiano, 2008; Turrigiano, 2011).

Different postsynaptic activities lead just to different speeds for the adaption of the weight vector length (see App. B.4). Thus, an adaption of the value α appears suitable to adjust the length of the weight vector. These synaptic changes can be referred to synaptic scaling, the change of synaptic efficiency (Abbott and Nelson, 2000; Turrigiano and Nelson, 2004; Turrigiano, 2011).

We implemented a form of activity dependent synaptic scaling as following. Each of our model neurons stabilizes its activity through a regulation of the length of its weights by an adaption of the parameter α_j (Eqn. 6.4). Which is adjusted so that each neuron uses a similar range for its activity r_j of roughly 0 to 1. This is achieved by an quadratic increase of α_j when the value of γ is exceeded, within the term H_j (Eqn. 6.5). γ denotes a soft upper bound for the activity range and is set to a value below one. Additionally, when strong excitatory currents exceed α_θ, than we also reduce the weight vector length. α_θ is set to a value above one. This should avoid a runaway increase of excitation because of a compensatory effect through inhibition, which keeps the activity in range while the currents increase. α_j decreases by the small constant ε. This constant decrease is adjusted in that way that α_j is reduced by the half of its value after a sufficient long learning period, when we do not regard any increase within this period. We have chosen a period length of 100000 stimulus presentations. Hence, we have two compensatory processes within our synaptic learning, the Oja normalization as rapid one and a very slow one by the adjustment of the allowed weight vector length (Zenke and Gerstner, 2017). The slow speed of this adjustment goes in line with the observed slow regulation processes in cell cultures (Turrigiano et al., 1998; Zenke and Gerstner, 2017).

$$\tau_\alpha \frac{d\alpha_j}{dt} = -\varepsilon + H_j + \left(\left(\sum_i w_{ij} r_i - \alpha_\theta \right)^+ \right)^2 \tag{6.4}$$

with

$$\alpha_j = (\alpha_j)^+$$

and

$$\tau_H \frac{dH_j}{dt} = \left(\left(r_j - \gamma \right)^+ \right)^2 - K - H_j \tag{6.5}$$

with

$$H_j = (H_j)^+$$

Both α_j and H_j are positive defined. H_j decays by its own value and a constant K. This constant allows H_j to decay to zero and shifts the threshold for an increase a bit higher. The value of α_θ is chosen high enough to allow neuronal activity in the range [0,1], the term should just prevent an ongoing increase of excitation in suboptimal model configurations. The time constant τ_H is chosen in the length of a stimulus presentation and τ_α is chosen so that the weight vector length regulation became a long term process. For parameter values see Table A.7.

6.2.4 Synaptic plasticity of inhibitory connections

For learning the inhibitory connections we employed anti-Hebbian learning, which decorrelates the neuronal responses and allows the network to learn the independent components of the input (Földiák, 1990; Falconbridge et al., 2006). Unlike the excitatory connections we used no calcium dependent learning, we are used the firing rates for determining the weight change. The design of the learning rule again based on our previous work (Wiltschut and Hamker, 2009; Teichmann et al., 2012). Teichmann et al. (2012) extended Wiltschut and Hamker (2009) by introducing thresholds on the pre- and postsynaptic activity and made the learning speed activity dependent. That is, pre- and postsynaptic neurons had to be active above a certain threshold to strengthen inhibition and the learning speed was decreased for low presynaptic activities, which causes faster weakening than strengthening of inhibition and prevents neurons from becoming permanent silent from too much inhibition.

In contrast to the previous rule, our new learning rule used a constant learning speed τ_c. Further, just a single threshold θ_c is applied on the postsynaptic activity in the Hebbian term (Eqn. 6.6). This avoids the issue of the previous learning rule that no learning occurred when the postsynaptic activity dropped below its threshold. Further, the small single postsynaptic threshold should prevent the total suppression of a neuron. This means that when the postsynaptic activity remains below the threshold, then the normalization reduces the inhibitory weight. Thus, the mechanism implements a lower limit on the neurons activity until which inhibition it can permanently suppressed. The learning rule reads

as follows.

$$\tau_c \frac{dc_{kj}}{dt} = r_k \cdot (r_j - \theta_c)^+ - \alpha_c \cdot r_j \cdot c_{kj} \tag{6.6}$$
$$c_{kj} = (c_{kj})^+$$

The weights have been positively defined, the inhibitory character of the rule comes from the sign in the activation function (see Sec. 5.2). The parameter α_c controls the strength of the normalization and with that the value which the weight c_{kj} can reach. Note, the inhibitory normalization term fundamentally differs from the Oja normalization (see previous section), which bases on the relation of the postsynaptic activity to the length of the weight vector. Here, inhibition is reducing this activity. In consequence, the weight increase would also saturate without normalization, when the postsynaptic activity goes to zero. The weight, with normalization, develops proportional to the covariance of the pre and postsynaptic activities. Thus, the weight represents the relation between the two components of the covariance (Eqn. 6.7). These are, the expectation value of the coactivity and the product of the expectation values of the activities. The covariance, as the Pearson correlation, implies the dependency between the activities. The relation of the covariance to the learning rule becomes easily clear when interpreting the parameter α_c as the expectation value for the presynaptic activity $E(X)$ multiplied with a factor. The factor just scales the value of the inhibitory weights. The expectation value of the coactivity $E(XY)$ and the expectation value of the postsynaptic activity $E(Y)$ is sampled over time.

$$cov(X,Y) = E[XY] - E[X]E[Y] \tag{6.7}$$

6.3 Measuring neuron and network properties

This section is the final evaluation section of our model. We will focus on the response properties of the neurons and neuron populations. Further, we evaluate the learned weights. We will demonstrate that a efficient neuronal code emerges through our plasticity mechanisms. We investigate the receptive field shapes of the neurons in the different model layers. Additionally, we test the V2 neurons on a novel hypotheses, which claims that they are sensitive to naturalistic textures. Furthermore, we describe the obtained weight distributions for different connections and we relate the learned weight strengths to the response correlation between the neurons. In addition, we relate the structural property,

whether neurons are connected or not, to the response correlation and compare the findings to experimental data. We also investigate how strong weights between strongly correlated neurons contribute to the total weight. Then, we show that the neurons in our network developed a gradually increasing translation invariance with layer depth. Finally, we examine the capability of the model for object recognition and compare the performances of the different layers.

6.3.1 Efficient coding

Sparseness An objective in computational neuroscience is to understand the goal of sensory coding. Forming sparse representations has been supposed as such a goal (Field, 1994). Sparse means, only a few neurons are active when encoding a stimulus. A sparse representation is attributed with several advantages, as a high representational and memory capacity, high fault tolerance, and several items can be simultaneously encoded (Földiák and Young, 1995). Further, it is intended to be energy efficient (Levy and Baxter, 1996). This supposed goal is supported by several computational models, which demonstrated that with sparseness as objective and natural scenes as input representations similar to V1 simple-cell receptive fields emerge (Olshausen and Field, 1996, 1997; Hoyer, 2004; Rehn and Sommer, 2007; Rozell et al., 2008; Weber and Triesch, 2008). On the other hand, it has been shown that such an objective is not needed when simulating the basic brain circuit, consisting of excitation and inhibition learned via Hebbian and anti-Hebbian plasticity (Földiák, 1990; Falconbridge et al., 2006; Wiltschut and Hamker, 2009; Zylberberg et al., 2011; King et al., 2013).

While many many experimental findings support the sparse coding hypothesis, several newer one suggest that maximizing sparseness is not actively intended by the cortex (Berkes et al., 2009; Willmore et al., 2011; Harris and Mrsic-Flogel, 2013). Berkes et al. (2009) examined the primary visual cortex of ferrets and rats and found that sparseness decreases with age and increases with the level of anesthesia, he reckons that high sparseness levels found in previous studies have been overestimated. (Harris and Mrsic-Flogel, 2013) described that, whereas in mouse visual cortex layer-2/3 neurons show sparse responses, neurons in layer L5 show rather dense responses. They conclude that the cortex may employ different coding strategies tailored for the respective targets. (Willmore et al., 2011) measured the sparseness in the visual cortex of macaque monkeys by a free-viewing visual search task. They showed that lifetime sparseness, which is typically measured instead of

the population sparseness, is not necessarily correlated to population sparseness. Further, they also remark doubts about the objective of the neuronal code and conclude from their results that it is not lifetime sparseness.

Lifetime sparseness can be explained as the sparseness of the responses of a single neuron on multiple stimuli. Whereas population sparseness denotes the sparseness of a neuron population encoding a single stimulus. Experimentally, it is difficult to measure population sparseness, as many neurons have to be measured in parallel (Willmore et al., 2011). Because of this lifetime sparseness is often measured in experimental studies and is assumed to behave similar to population sparseness (Willmore et al., 2011). This interchangeability is not given in general and has been criticized (Lehky et al., 2005; Berkes et al., 2009; Willmore et al., 2011; Spanne and Jörntell, 2015). However, lifetime and population sparseness can be similar for independent responses (Berkes et al., 2009; Willmore et al., 2011), which are identically distributed (Willmore et al., 2011). Note, that population sparseness is important as it measures how efficient stimuli are encoded.

Several measurements for sparseness have been established. We will briefly introduce three of them. First the kurtosis (Eqn. 6.8), which is the fourth moment, the tailedness, of a distribution (Field, 1994; Olshausen and Field, 2004; Lehky et al., 2005).

$$S_K = \frac{1}{n}\sum_{i=1}^{n}\frac{(r_i - \bar{r})^4}{\sigma^4} - 3 \tag{6.8}$$

Where n is the amount of responses, r_i the ith response, $\bar{r}$ the mean response, and σ is the standard deviation of the responses. High values indicate a super-Gaussian response distribution and, thus, high sparseness. Whereas negative values indicate a less sparse sub-Gaussian distribution. Because of its fourth power the measure is very sensitive to outliers (Olshausen and Field, 2004; Lehky et al., 2005). Examples for the measure can be found in Tolhurst et al. (2009, Fig. 3). We do not use it in our work.

The second measure, is the fraction of the squared mean activity to the mean squared activity, called Treves-Rolls or Rolls-Tovee sparseness (Treves and Rolls, 1991; Rolls and Tovee, 1995). The measure has been slightly modified by Vinje (2000) to give values in the range [0,1]. Where zero indicates dense codes and one an one-out-of-n code. It has been further modified (Eqn. 6.9) to be also valid for negative response values (Berkes et al.,

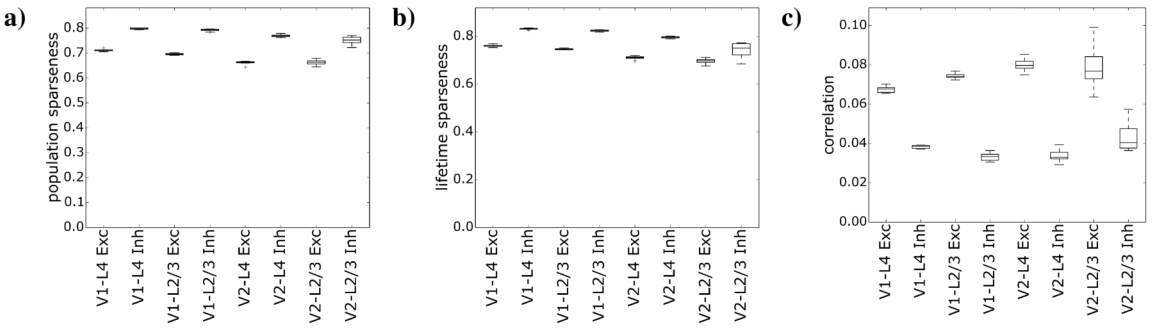

FIGURE 6.1: Sparseness and correlations of the neurons in every population. The sparseness and correlations between the responses of all neurons on 100000 natural scene stimuli have been measured with the Rolls-Tovee sparseness measure and linear correlations. a) Shows the box plot for the sparseness values each neuron population had on the stimuli, called population sparseness. b) The sparsenesses of the response of every single neuron in every population, called lifetime sparseness. c) The linear correlations of the responses between all pairs of neurons within a each population. The sparseness slightly decreases in deeper layers. Inhibitory populations develop higher sparseness than excitatory one. The correlations behave similarly, inhibitory populations show less correlations than excitatory populations. In general correlation values are very low, while the sparseness values are less sparse than assumed in sparse coding models.

2009; Willmore and Tolhurst, 2001).

$$S_R = 1 - \frac{\frac{(\sum_{i=1}^{n} |r_i|/n)^2}{\sum_{i=1}^{n} r^2/n}}{1 - \frac{1}{n}} \tag{6.9}$$

Again n denotes the amount of responses and r_i a single response. The equation is closely related to the formulation of the variance (see Lehky et al., 2005). Further, it is sensitive to the mean and variance of the distribution (Lehky et al., 2005). The values obtained with that measure are found to be closely related to the kurtosis (Willmore and Tolhurst, 2001). Examples for the measure can also be found in Tolhurst et al. (2009, Fig. 3).

The third measure, is the relation of the L1-norm to the L2-norm (Eqn. 6.10), called Hoyer sparseness (Hoyer, 2004). It gives values in the range [0,1], similarly to the Rolls-Tovee measure.

$$S_H = \frac{\sqrt{n} - \frac{\sum_{i=1}^{n} |r_i|}{\sqrt{\sum_{i=1}^{n} r_i^2}}}{\sqrt{n} - 1} \tag{6.10}$$

Again n denotes the amount of responses and r_i a single response. Examples for the measure can be found in Hoyer (2004, Fig. 2) and Wiltschut and Hamker (2009, Fig. 3).

We evaluated the different sparseness values for any neuron population for 10 indepen-

dent model runs and report the mean values and the box plots of these runs. For the population sparseness (Fig. 6.1a) we obtained values (averaged over all populations) around 0.73 with the Rolls-Tovee measure. Similarly, the lifetime sparseness (Fig. 6.1b) was measured with 0.76. The close values are presumably caused by the high independence of our neurons. Interestingly, inhibitory neurons show a higher population sparseness than excitatory one (0.78 to 0.68) as well as lifetime sparseness (0.79 to 0.72). This might be explained by the connection structure. Inhibitory neuron populations inhibit itself and their efferent excitatory population. Thus, the decorrelation effect might be stronger. The opposite has been found in the mouse cortex (Harris and Mrsic-Flogel, 2013). There, inhibitory interneurons have been described to use a rather dense code, caused by an unspecific connectivity and non-sensory inputs (Harris and Mrsic-Flogel, 2013). Since our model neurons have no non-sensory inputs our results might differ from physiological studies of freely behaving animals. However, in studies using anesthesia, which is known to reduce background inputs, commonly higher sparseness levels are found. Consider, that the inhibitory circuit of our network is undercomplex in comparison to the one found in cortex, where at least three different interneuron types are described to form a complex network of inhibition and disinhibition (Harris and Mrsic-Flogel, 2013). However, our network, as all networks employing an learning rules (which are Hebbian associative), develops specific connections related to the correlations between the neuron activities (e.g. King et al., 2013) and is not just averaging the surrounding activities to determine the strength of inhibition. The resulting tuned receptive fields have been also found in Cats V1-L4 for different interneuron types (Hirsch et al., 2003). Thus, without non-sensory inputs, the low complexity of the model inhibitory circuit, and tuned inhibitory neurons very low sparseness levels seem unlikely for our inhibitory interneuron populations. Further, we observed that with increasing layer depth the sparseness values decreased. A potential reason for that could be the decrease of neurons in the populations, changing the inhibition between the remaining neurons when we used the same parameters for inhibitory plasticity. However, a similar behavior has been found in the visual cortex, where the lifetime sparseness slightly decreases from area V1 to area V4 (Willmore et al., 2011). Albeit, the authors of the study found the evidence for the effect not convincing and interpreted their results as side effect of the stimulus protocol, which differed between V1/V2 and V4.

The sparseness values obtained with the Hoyer measure behave similarly, but with lower values (Fig. D.1). We obtained a mean population sparseness of 0.52 and a mean life-

time sparseness of also 0.52. This value is much lower than that which Hoyer (2004) has used as target sparseness (0.85) to obtain simple-cell like receptive fields. Further, an earlier simple-cell modeling study of our lab, using related learning rules, obtained population sparsenesses between 0.72 and 0.92 for different parameterizations (Wiltschut and Hamker, 2009). An other simple-cell modeling study with similar learning rules to ours obtained also higher sparseness levels than we do (King et al., 2013). They measured a population and lifetime sparseness of 0.96 with the Rolls-Tovee measure. We attribute the difference between the model sparsenesses to the configuration of the inhibitory mechanism. In (Wiltschut and Hamker, 2009) the inhibition was configured stronger, because of a non-liniarity on the inhibitory signal and the tuning of the parameters to achieve receptive fields which best match physiological data. In King et al. (2013) the inhibitory weights can develop to higher values for a certain correlation than in our parametrization. This is because he aligns the weights without a controlling factor to the average activities. Whereas, we have chosen our parameter α_c much higher than the average activity, which applies a factor below one on the weight height and causes lower weights (see Sec. 6.2.4). Further, the short time trace for estimating the average values in King et al. (2013) probably underestimates the real average activity of the neurons and leads to higher weights than the equation suggests and with that to stronger inhibition and higher sparseness.

Correlation Barlow (1961) formulates the hypothesis that a goal of the sensory system might be redundancy reduction. This appears quite likely as a biological system can not waste much resources. However, Barlow realizes that Shannon's formulation of redundancy is not an adequate measure for redundancy in the brain: "Because the brain uses information in different ways from those common in communication engineering" Barlow (2001). The brain needs a representation where it can discover regularities, i.e. items should be easily accessible, and statistical dependencies should be low (Barlow, 2001). The basic Hebbian term is exploiting the dependencies between pre- and postsynaptic activity. Hence, each neuronal layer should learn a compact representation of these easy detectable dependencies and forming a structure which has not the same dependencies anymore.

We measured the pairwise linear correlations (Eqn. 6.11), also known as Pearson correlation, to get insights in the dependencies within each neuron population. The correlations are intended to be low in sparse codes. When the linear correlation between the responses

of two neurons is zero, the neurons have no linear dependence. Which indicates, but not grants, a substantial degree of independence. Independence by itself is an important property for an efficient neuronal code. When all pairs of neurons are independent the neuronal code should have no redundancy (Olshausen and Lewicki, 2013).

$$corr(X,Y) = \frac{E[(X-\bar{X})(Y-\bar{Y})]}{\sigma_x \sigma_y} \tag{6.11}$$

X and Y are the responses of two neurons on multiple stimuli and $\bar{X}$, $\bar{Y}$ are the respective mean responses. σ_x and σ_y denote the standard deviations of the responses.

In experimental studies (Smith and Kohn, 2008; Ecker et al., 2010) as well as in computational models (Wiltschut and Hamker, 2009; Zylberberg et al., 2011; King et al., 2013) correlations have been found as very low. We also obtained very low correlations of around 0.056 averaged over all populations and 10 model runs. As indicated by the sparseness values, the correlations within inhibitory populations are lower than within excitatory populations (Fig. 6.1c). We attributed this to the stronger effect of inhibition on these populations. In contrast to the sparseness data we did not observe a consistent trend of increasing correlations with increasing depth of the layers. While the correlations in the excitatory populations seem to increase up to V2-L4, the correlations in the inhibitory populations seem to follow the opposite trend. For both, inhibitory and excitatory populations, the assumed trend is broken in V2-L2/3 and the large deviation between the different model runs obfuscate the results. Anyway, a trend would be very weak on a very high level of decorrelation. This high level of decorrelation was also found in other models (Wiltschut and Hamker, 2009; Zylberberg et al., 2011; King et al., 2013). Also physiological studies found very low correlations. For instance $\leq$0.02 in the primary visual cortex of awake macaques (Ecker et al., 2010). Smith and Kohn (2008) found low correlations, similar to our results, for distant neurons in the primary visual cortex of macaques and higher correlations for proximate neurons. He obtained an average value of 0.176 over all measured neurons. The trend of decreasing correlations with increasing cortical distance is consistent to our data of the neurons in V1-L4 (Fig. D.2). This appeared obvious to us as objects are spatially compact and the basic features encoded by V1 are unlikely to show high correlations to more distant feature in the visual space.

6.3.2 Receptive field shapes

The receptive fields of the neurons in the primary visual cortex (V1) can be distinguished into mainly two groups, called simple and complex (Hubel and Wiesel, 1962). Simple receptive fields can be described by a map of spots of light with regions where light has an inhibitory or excitatory effect on the neuron responses. Whereas for complex receptive fields the description by such a map is not possible (Hubel and Wiesel, 1962). The receptive field shapes of simple-cells have been found to be well described by Gabor functions (Jones and Palmer, 1987). Complex receptive fields in V1 are similar to simple receptive fields, but are slightly invariant to the position of the stimulus (Hubel and Wiesel, 1962; De Valois et al., 1982; Adelson and Bergen, 1985; Carandini et al., 2005). Simple receptive fields are the predominant type in V1-L4 and complex receptive fields in V1-L2/3 (Hubel and Wiesel, 1962; Ringach et al., 2002). Following the taxonomy of simple and complex all receptive fields in deeper layers are complex, thus, this taxonomy is only used for V1 neurons. The responses of neurons in the next area of the ventral pathway, the extra striate cortex (V2), are more difficult to understand. A natural hypothesis is that they respond to combinations of features represented in V1-L2/3, the layer projecting to V2-L4. V2 neurons have been found experimentally to respond to simple grating, bar, or sinusoidal stimuli as well as contours and textural characteristics, whereas the ability of discriminating contours seems higher (Hegdé and Van Essen, 2000). These different kind of represented shapes evoke well separable population responses (Hegdé and Van Essen, 2003). The neurons have been further described as selective to angle stimuli (corners) (Ito and Komatsu, 2004; Anzai et al., 2007). However, the responses to the optimal angle was found to be very similar to the response to an optimal line stimulus (Ito and Komatsu, 2004). An unpublished study of us, with one learning layer on top of a Gabor energy model (spatiotemporal energy model Adelson and Bergen, 1985), suggested a receptive field preference for elongated edges over contours. This early result goes in line with the findings that the majority of V2 neurons have receptive fields with subregions responding mainly to similar orientations Anzai et al. (2007). Another study found that nearly the half of the V2 neurons are tuned similar to V1, when stimulated with natural images (Willmore et al., 2010). To investigate the differences between V1 and V2, Freeman et al. (2013) compared the responses of both areas, using naturalistic stimuli and the same stimuli with randomized phase structure, which removes higher-order correlations from the stimuli. V1 neurons respond, as expected, similarly on both stimuli, whereas V2 neurons respond

stronger to the naturalistic ones. Thus, V2 neurons are sensitive to higher-order correlations of naturalistic stimuli. Nonetheless, it remains still unclear what kind of receptive fields the neurons in V2 have. The evidences can be summarized to they are sensitive to basic shape elements, similar to V1, or some which can be formed from V1 representations. With these elements they are able to detect the basic elements of visual scenes as textures, corners, edges, while having a higher sensitivity to naturalistic arrangements.

V1-L4 feedforward weight matrices

To gain insights in the learnings of the neurons in our first network layer (V1-L4), we visualized the feedforward weight matrices of the neurons. Since the feedforward weights are the only source of input to the network, they should dominate the functioning of the neurons. The structure of the weights should appear Gabor like, with a sufficient diversity of position, extend (size), orientation, and frequency properties. We visualized the weight matrix of the first 100 neurons. The resulting images show to which LGN neurons the cells are connected. Because of the On-Off structure of LGN, we show the connections to the on- and off-neurons in LGN first separated, as they could potentially overlap. Further, as the size of both parts could differ and, thus, the weight strengths could differ, we plotted the data with different normalizations on the parts, resulting in four different plots. 1) We visualized just the connections to the LGN on-neurons. 2) To the LGN off-neurons. 3) We subtracted the off part from the on part and used the same colors as in 1 and 2 for the positive and negative part. For weights, having the same weight strength, the same brightness is used. 4) We do the same as in 3, but we use the full color range for each part. That is, the strongest weights in each part get the darkest colors.

Each tile in an image shows the receptive field of an individual neuron. The tiles are separated by a black border. The weight matrix of each neuron is individually normalized to use the full color range. Bold colors denote strong weights and bright weak weights. White denotes a weight strength of zero or no connection. For weights to the on-neurons in LGN, we used red color and to the off-neurons blue (cf. Cossell et al., 2015).

Excitatory neurons We found receptive field shapes of various sizes and orientations (Fig. 6.2). The structural plasticity allowed the neurons to grow larger than their initial connectivity of 12 by 12 (see Sec. 2.2), thus, several neurons developed large receptive fields (cf. Sec. 4.3.3). However, other neurons developed blob-like receptive fields.

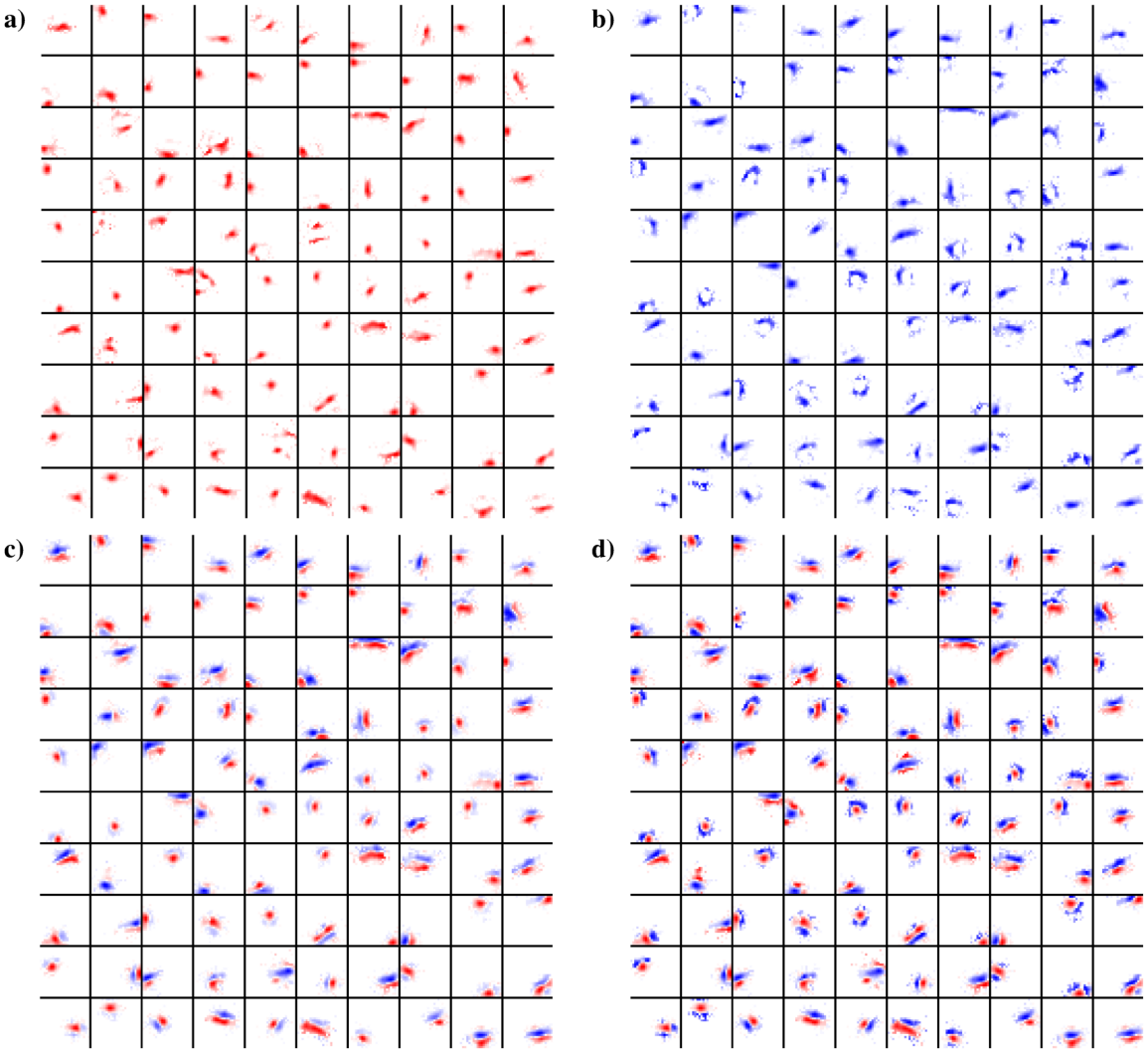

FIGURE 6.2: Weight matrices for the connection of the LGN neurons to the first 100 V1-L4 *excitatory* neurons. Each tile shows the receptive field of an individual neuron. The weight matrix of each neuron is individually normalized to use the full color range. Bold colors denote strong weights and bright weak weights. White denotes a weight strength of zero or no connection. **a)** Shows the weights emanate from the LGN on-center neurons. **b)** Weights from the off-center neurons. **c)** The difference between the on- and off-weights, where blue denotes negative values, related to the off-weights and red positive ones, related to the on-weights. Here, the neuron weights are scaled to $\pm$max value of on- and off-weights. **d)** Similar to c, but the weights are scaled to use the full color range for each neuron.

These have been found in physiological studies (Ringach, 2002), but have been missed in early computational models, such as independent component analysis (Bell and Sejnowski, 1997; van Hateren and van der Schaaf, 1998; van Hateren and Ruderman, 1998) or sparse coding (Olshausen and Field, 1996, 1997).

Rehn and Sommer (2007) again developed a model with a hard sparseness constraint (how many neurons are active) and showed that with this receptive field shapes, found in macaque, can be better predicted than by soft sparseness models (constrains the overall activity). Their model (SSC) was able to predict the blob-like receptive fields as well as receptive fields with multiple subfields in a comparable amount to macaque monkey data, while the earlier soft sparseness approach (Sparsenet) of (Olshausen and Field, 1996) just predicted Gabor like receptive fields (Fig. 6.4a). For comparison, we show the feedforward weight matrices of our V1-L4 excitatory model neurons (Fig. 6.4b). We ordered the receptive fields of the neurons by the spatial frequency multiplied with the spatial extend in the y-dimension (n_y, cf. Sec. 4.3.3). We plotted 98 neurons in ascending order, in equal steps over the criteria. Note, each tile of our model neurons has 24 by 24 pixels, whereas the other data have 16 by 16. We found a comparable amount of blob-like and non orientation selective receptive fields as in the macaque data, but we found no significant amount of receptive fields with more than three subfields. When regarding the Gabor fits and using other ordering criteria, as large x-extends, we see several receptive fields with more than three subfields. But when comparing the weight matrix not more than three subfields are present. We are using a similar Gabor fit function (Sec. B.2) to (Ringach, 2002), thus, we speculated that the high amount of subfields in the data of Ringach might be an artifact of the used Gabor fit method.

Inhibitory neurons The shapes of the feedforward weights of the inhibitory neurons also have various sizes and orientations (Fig. 6.3). More units than in the excitatory population developed blob-like receptive fields. This finding is similar to a physiological study, in the cat V1 layer-4, where also orientation selective and non-selective interneurons have been found (Hirsch et al., 2003). In a computational model, resembling V1-L4 with a circuit of excitatory and inhibitory neurons and learning rules related to ours, also orientation tuned receptive fields of the inhibitory interneurons developed, although with lower spatial frequencies in comparison to their excitatory receptive fields King et al. (2013, Fig. 3). In contrast to (King et al., 2013), our inhibitory interneurons get feedforward input from

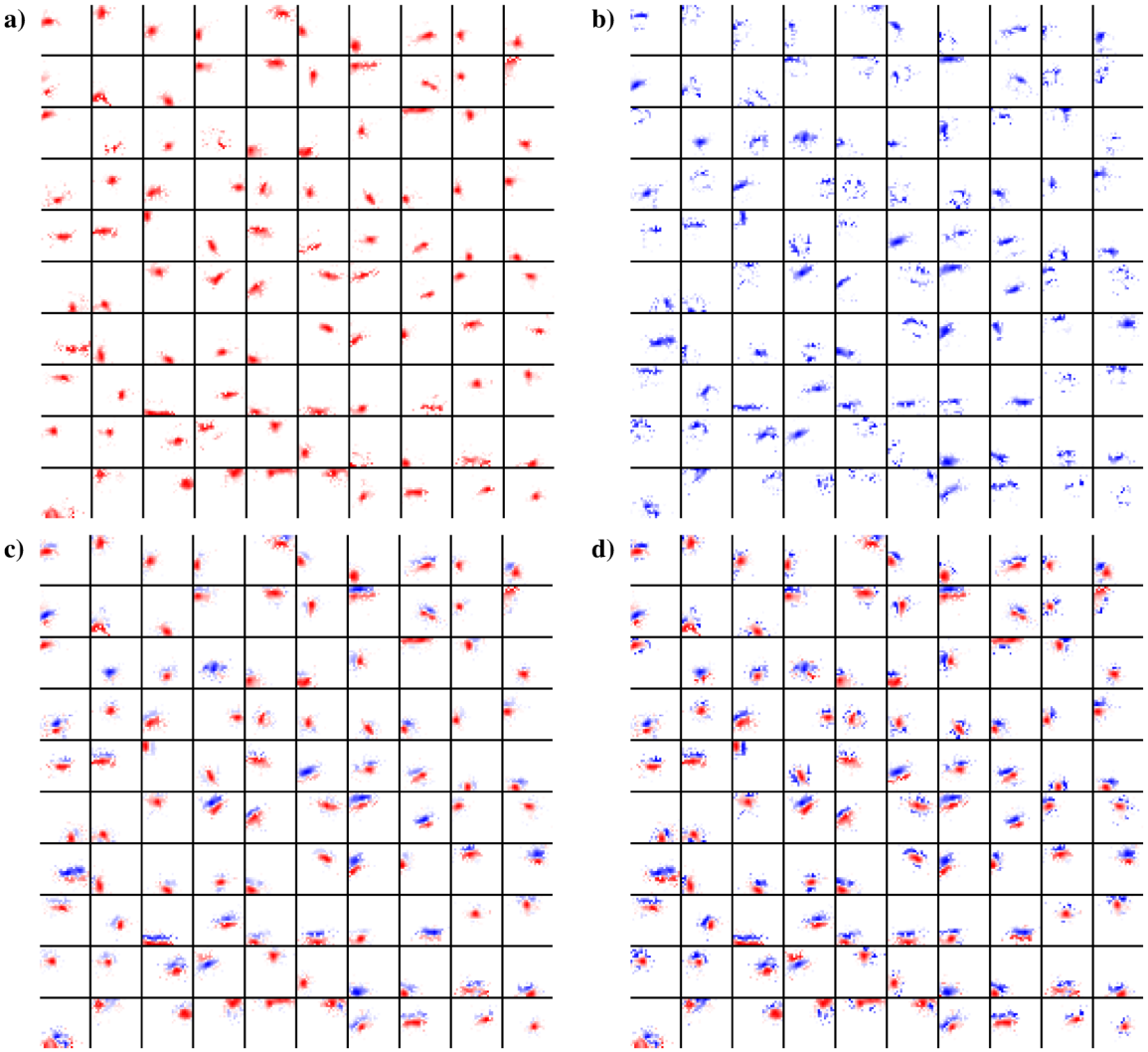

FIGURE 6.3: Weight matrices for the connection of the LGN neurons to the first 100 V1-L4 *inhibitory* neurons. Each tile shows the receptive field of an individual neuron. The weight matrix of each neuron is individually normalized to use the full color range. Bold colors denote strong weights and bright weak weights. White denotes a weight strength of zero or no connection. **a)** Shows the weights emanate from the LGN on-center neurons. **b)** Weights from the off-center neurons. **c)** The difference between the on- and off-weights, where blue denotes negative values, related to the off-weights and red positive ones, related to the on-weights. Here, the neuron weights are scaled to $\pm$max value of on- and off-weights. **d)** Similar to c, but the weights are scaled to use the full color range for each neuron.

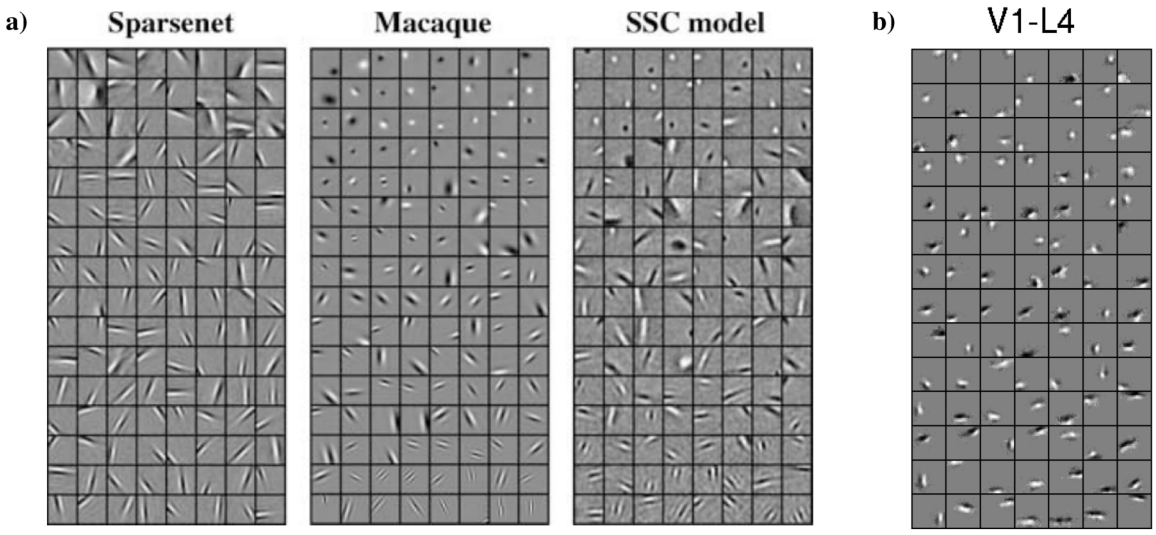

FIGURE 6.4: Comparison of receptive fields to shapes obtained by other models and physiological studies. . **a)** "Receptive fields from the efficient coding models and from recordings in monkey V1. The models were trained on 16 × 16 patches of natural input. Each panel shows 128 randomly selected cells, ordered with respect to shape. Experimental results are shown as Gabor fits (data courtesy of D. Ringach). Scale differences due to distance from the fovea were corrected for" (Rehn and Sommer, 2007, Fig. 5). **b)** Feedforward weights of 98 excitatory neurons in V1-L4 of our model, sorted by the spatial frequency multiplied with the spatial extend (y). Ploted in the same colormap as a) where bright denotes weights to on-regions and dark off-regions, grey denotes zero weights. Note, each tile shows 24 by 24 pixel.

LGN, as the excitatory neurons. This allows them to learn similar feedforward weights. However, they visually appear less Gabor like than the one of the excitatory neurons. This might be caused by the second excitatory drive from the excitatory neurons, which makes the feedforward connections from LGN less relevant for their behavior. Thus, just regarding the feedforward weights might be not sufficient when evaluating inhibitory interneurons.

V1-L4 reverse correlation

To gain more insights in the behavior of the excitatory and inhibitory neurons, we calculated the receptive fields from the responses of the neurons with reverse correlation. The reverse correlation maps the responses on the input linearly into the visual space (Ringach and Shapley, 2004; Sharpee, 2013). This gives us an estimation of the visual shapes to which the neurons are sensitive. However, this approach is limited to simple receptive field structures. Higher order neurons often have largely overlapping receptive field components

and respond to complex structures, so that sufficient stimuli are difficult to generate under the condition to have an adequate statistic of the input, as white noise has (Ringach and Shapley, 2004).

To calculate the reverse correlation, we stimulated the network with uniformly distributed white noise and recorded the responses. Then, we weighted any stimulus with the response of the respective neuron and accumulated the values. As result we obtained a response map in the input space with on- and off-regions. Similarly to what we have done for the feedforward weights, we subtracted them from each other and obtained one image, where we colored the on-regions red and the off-regions blue. We visualized again the first 100 neurons from both populations of V1-L4 (Fig. 6.5).

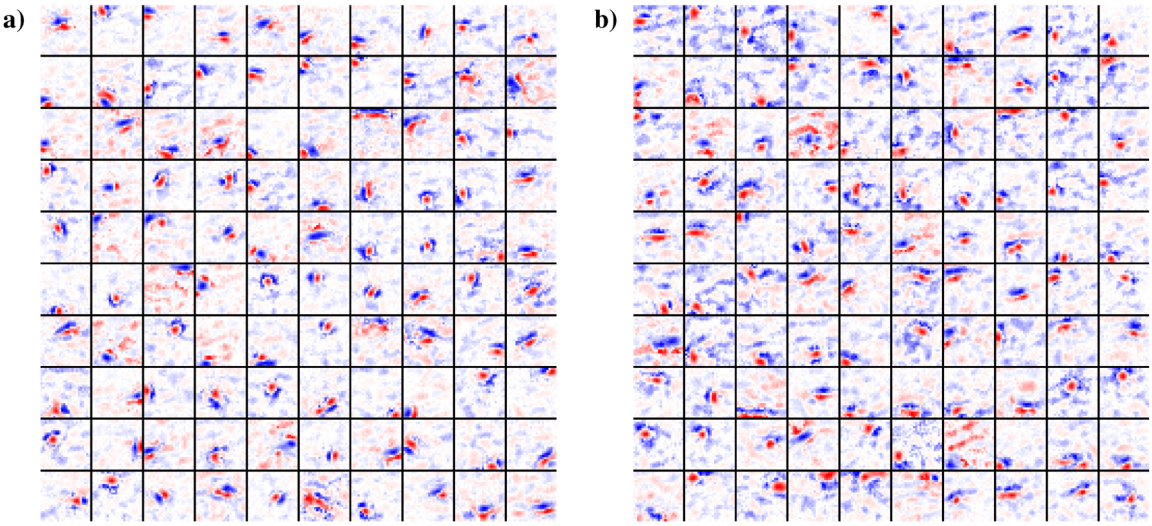

FIGURE 6.5: Receptive fields of the first 100 V1-L4 excitatory and inhibitory neurons measured via reverse correlation. Each tile shows the receptive field of an individual neuron. The values for on-center and off-center LGN neurons are subtracted. The resulting matrix for each neuron is individually normalized to use the full color range. Bold colors denote strong and bright weak influence on the activity. Blue denotes off-neurons and red on-neurons of LGN. a) Shows V1-L4 excitatory neurons and b) inhibitory neurons.

The most receptive fields of the excitatory neurons appear similar to the shapes obtained from the feedforward weights (Fig. 6.5a). Also the receptive fields of the most inhibitory interneurons appear similar (Fig. 6.5b). Although, larger and stronger regions of other on- and off-regions are visible. Partially, they are an artifact of the reverse correlation method, emphasized by the region wise normalization of the data for visualization. Similar effects also appear in physiological recordings (e.g. Cossell et al., 2015; Ringach, 2002).

Nevertheless, some of the structures reflect inputs the inhibitory neurons receive from the excitatory neurons or from other inhibitory neurons. Which might indicate sensitivity to more complex structures than Gabor like stimuli.

V1-L2/3 weight projections

Unfortunately, a linear mapping of the weights in higher layers is not simply possible. Subfields are likely to overlap and no structured maps of the subfields can be plotted anymore (cf. Hubel and Wiesel, 1962). In computational models the connection strengths between all neurons are available, thus, the second layer neurons can be described by a the set of first layer neuron receptive fields having strong weights (Hyvärinen and Oja, 2000).

We plotted for each excitatory neuron in V1-L2/3 30 feedforward weight matrices of strongly connected excitatory neurons from V1-L4. We ordered the receptive fields by the connection strength and weighted each field with that strength, so that fields with a weak connection appear weaker. Again, we used red colors to denote on-regions and blue for off-regions. The neurons in V1-L2/3 should be majority so called complex-cells, i.e. selective to a single orientation and fields at slightly different positions. We show eight selected neurons with different receptive fields.

The most neurons are selective to V1-L4 neurons selective to a single feature at neighboring positions (Fig. 6.6). Also several neurons have also learned connections to similar V1-L4 neurons with blob-like receptive fields. Some neurons developed connections to two distinct groups of V1-L4 neurons with different oriented receptive fields. However, it remains unclear if these connections make the neuron sensitive for combinations of edges (corners) in the input. A fraction of V1-L2/3 neurons with more than one orientation preference would indicate the possibility of a continuous increase in complexity over the cortical hierarchic instead of a strict hierarchy, where in each layer a fraction of neurons is doing similar things to neurons in lower or higher layers.

V2

As mentioned above, subfields in deeper layers are largely overlapping and the neurons are poorly responding on white noise input (Ringach and Shapley, 2004). However, in computational models we can easily access all network properties, as weight strengths and responses of all neurons, which would be tough in physiological studies. With that we can map the weights of each neuron back into the input space. Unfortunately, this leads to a

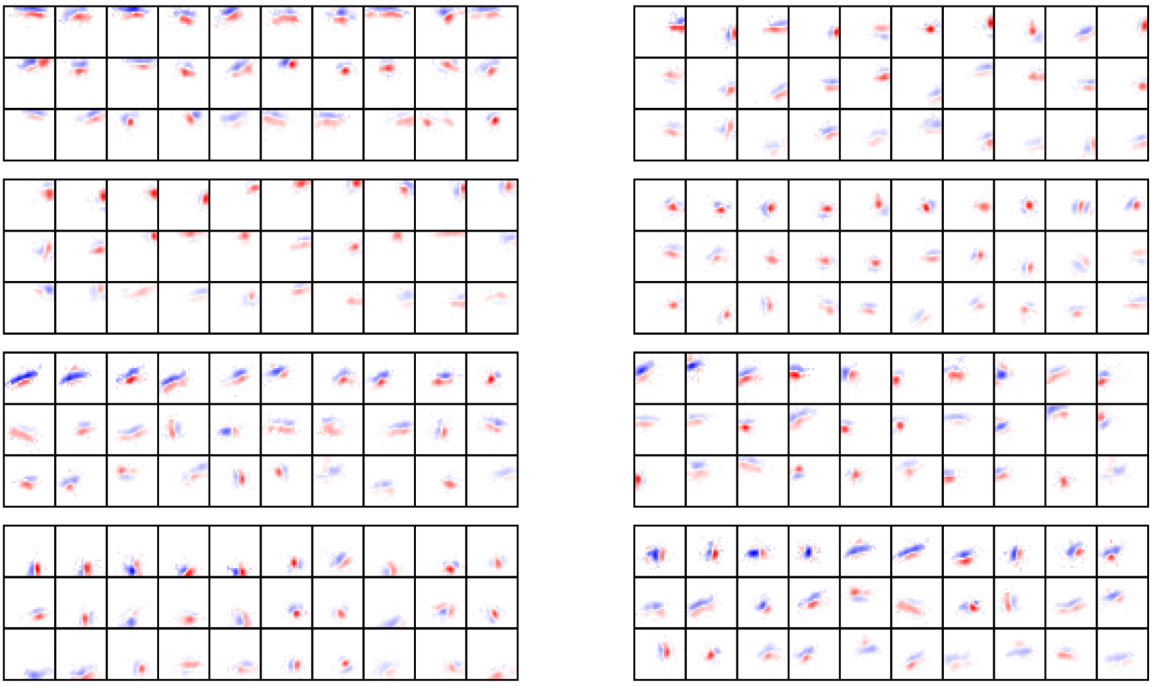

FIGURE 6.6: Receptive fields eight selected V1-L2/3 excitatory neurons. Each box shows the receptive field of an individual neuron, visualized by plotting the weight matrices of the afferent excitatory neurons of V1-L4 (tiles). The weight matrices of the V1-L4 are weighted by the weight strength to the regarded V1-L2/3 neuron. Bold colors denote strong and bright weak influence on the activity. Blue denotes weights to off-neurons and red to on-neurons in LGN.

"soup" of multiple paths ending at the same pixels, so that the sum of all paths reveals no specific pattern for which a neuron is responsive. When subtracting the resulting on- and off-regions, we obtained just a shape formed through the subtraction of two overlapping Gaussians with no meaning for the neuron's sensitivity. An alternative technique to describe what a neuron does is to search for their optimal stimulus (Le et al., 2012). However, we used natural scene stimuli and each neuron responds just to parts of a stimulus patch, so that the optimal stimulus could be difficult to interpret and misleading when regarding the optimal natural scene patch. Thus, we combined both methods to highlight the structures of the optimal stimulus which contributed to the neurons response.

Therefore, we searched for the optimal stimulus of each neuron by presenting natural scenes until the neuron is highly activated (response close to 1). When a stimulus was found, we recorded the activities of all neurons in the network and stored the stimulus. We repeated this until we found for all network neuron a stimulus. With increasing number of presented stimuli, we relaxed stepwise the condition on the activity for accepting a stim-

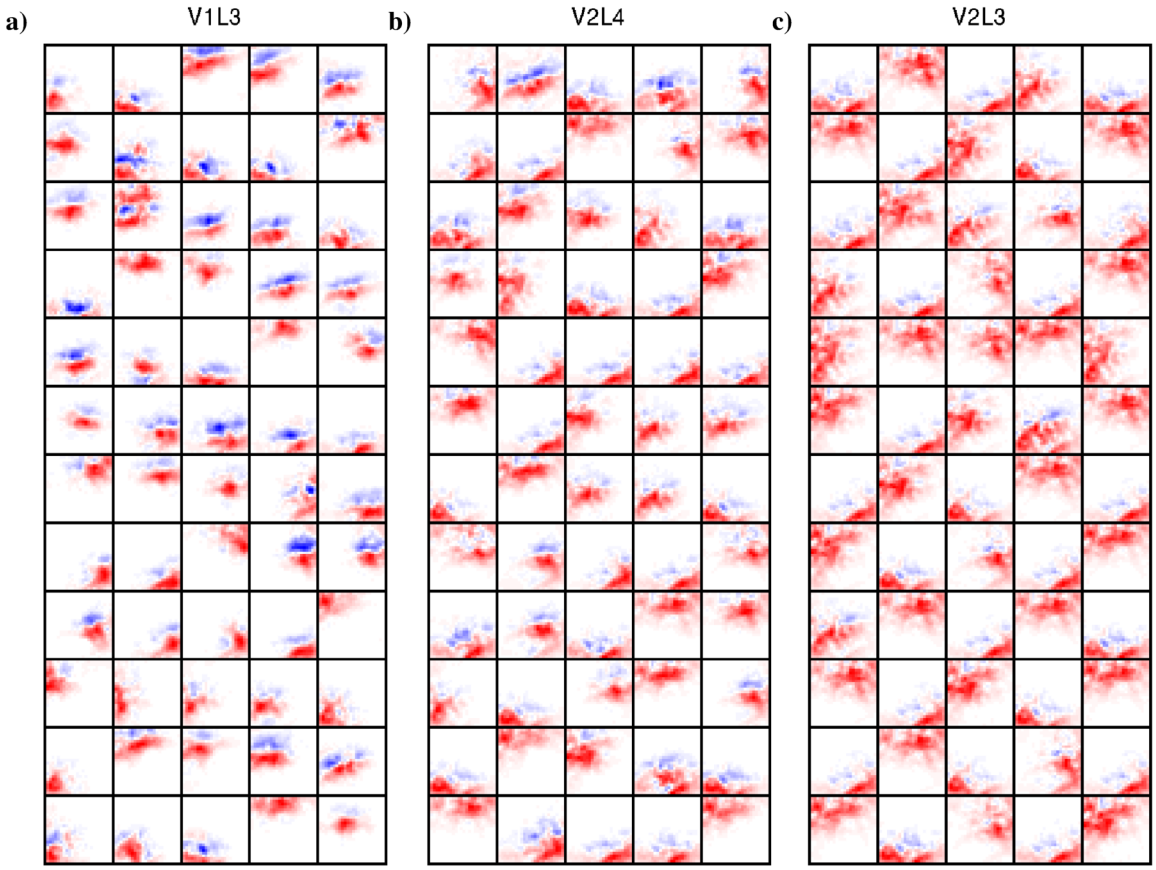

FIGURE 6.7: Back projection of the weight matrices into the input space for the excitatory neurons of different layers. Each tile shows the weights to the input of an individual neuron. The values for on-center and off-center LGN neurons are subtracted. The resulting matrix for each neuron is individually normalized to use the full color range. Bold colors denote strong and bright weak influence on the activity. Blue denotes off-neurons and red on-neurons of LGN. **a)** Shows 60 V1-L2/3 excitatory neurons, **b)** 60 excitatory V2-L4 neurons, and **C)** 60 excitatory V2-L2/3 neurons.

ulus. In total we presented about 100000 natural scene patches until an optimal stimulus was found for each neuron. Following, we projected the excitatory feedforward weights of each neuron back through the hierarchy into the input space. We weighted each connection with the neuron activities the neurons had on the optimal stimulus of the regarded neuron. Finally, we weighted the projection with the input image.

To illustrate the the degree of overlapping subfields, we visualized the back projected weights for 60 excitatory neurons of each layer in V2 (without weighting them with the responses on the optimal stimulus). For comparison we visualized 60 excitatory neurons from V1-L2/3. Whereas, the images indicate Gabor like shapes in V1-L2/3 they become

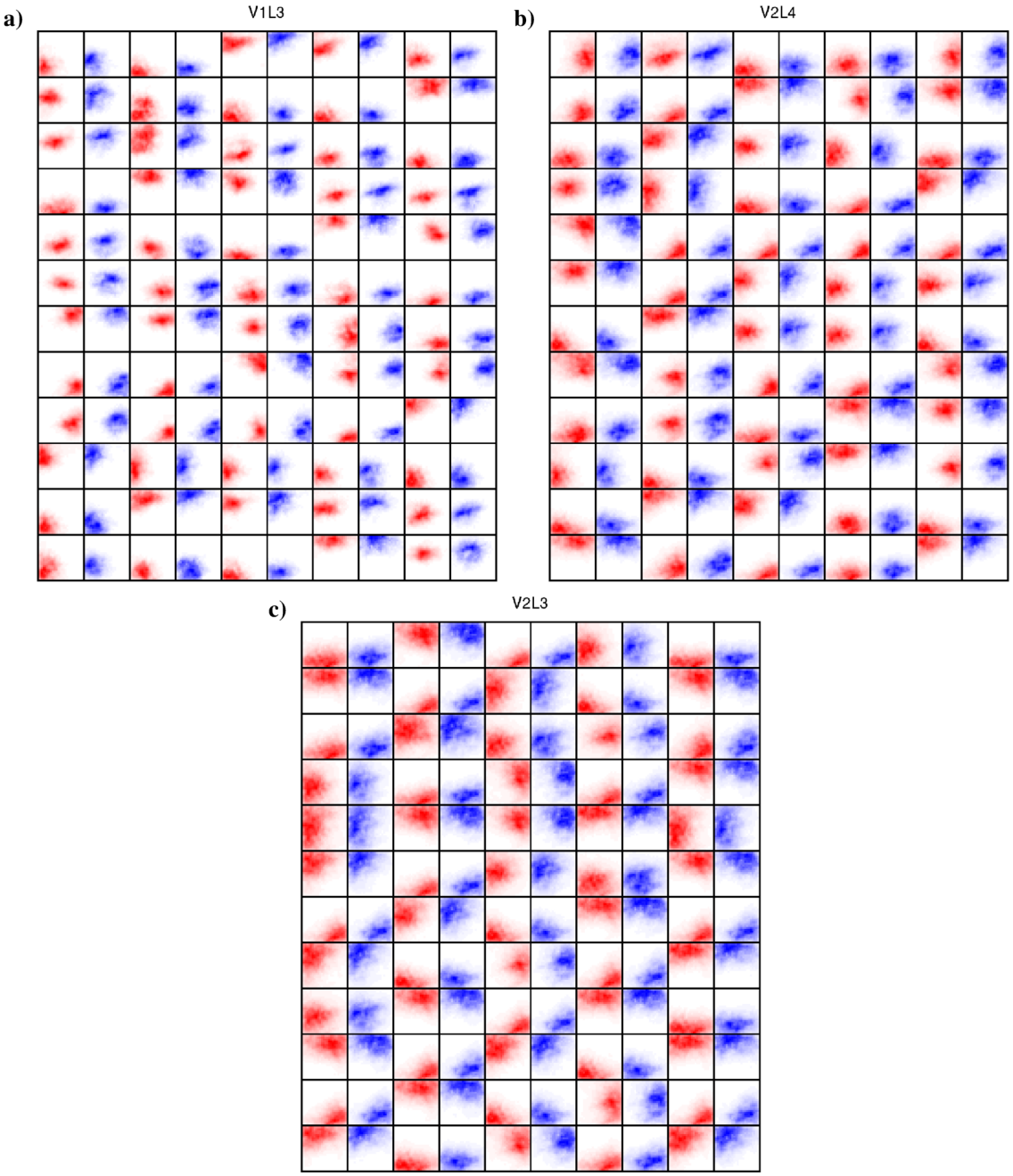

FIGURE 6.8: Back projection of the weight matrices into the input space for the excitatory neurons of different layers, with separated on- and off-planes. Each two neighboring tiles show the weights to the input of an individual neuron. The values for on-center and off-center LGN neurons are shown beside each other. The weight matrix for each neuron is individually normalized to use the full color range. Bold colors denote strong and bright weak influence on the activity. Blue denotes off-neurons and red on-neurons of LGN. **a)** Shows 60 V1-L2/3 excitatory neurons, **b)** 60 excitatory V2-L4 neurons, and **C)** 60 excitatory V2-L2/3 neurons.

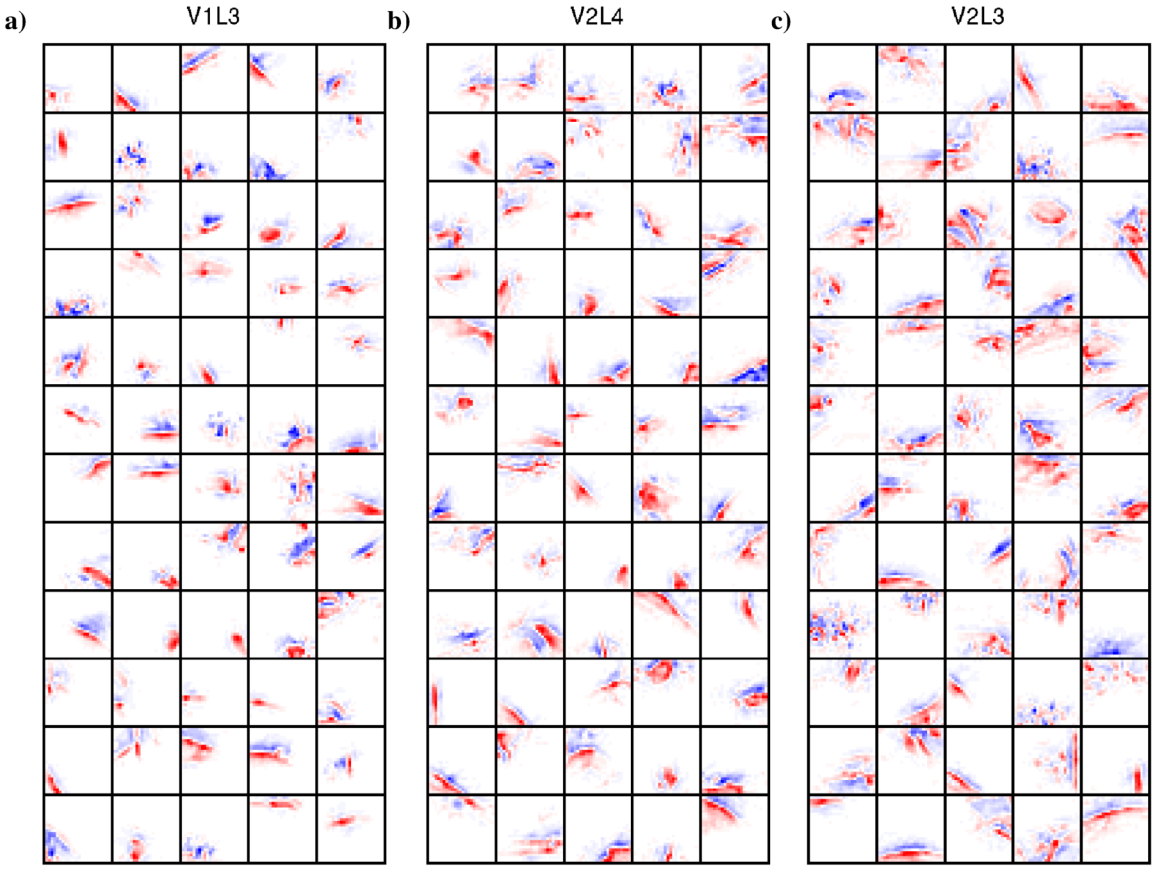

FIGURE 6.9: Back projection of the weight matrices into the input space for the excitatory neurons of different layers, weighted with the network activity on the optimal stimuli of the neurons. Each tile shows the weights to the input of an individual neuron, the back projected weights have been multiplied with the activities of the neurons. Thus, just image parts driving the neuron are shown. The values for on-center and off-center LGN neurons are subtracted. The resulting matrix for each neuron is individually normalized to use the full color range. Bold colors denote strong and bright weak influence on the activity. Blue denotes off-neurons and red on-neurons of LGN. **a)** Shows 60 V1-L2/3 excitatory neurons, **b)** 60 excitatory V2-L4 neurons, and **C)** 60 excitatory V2-L2/3 neurons.

more and more diffuse in V2 (Fig. 6.7). Then, regarding the two subfields separate, it can be seen that also in V1-L2/3 the fields have large overlap which becomes larger with the increase of the receptive field size in deeper layers (Fig. 6.8). The, especially in V1-L2/3, appearing feature of an oriented Gabor would be a likely feature, but it is just the result of two Gaussian shaped subfields with different centers and do not indicate an orientation feature. The intermediate neurons, with weights to this image parts, could encode any feature.

We weighted the back projected weights with the activities on the optimal stimulus of the regarded neuron. We observed edge like structures which appear more complex in deeper layers (Fig. 6.9). This might have its reason in the increasing receptive field size in deeper layers. However, we regarded just input parts evoking a response in neurons connected to the regarded neuron. This means, just image parts related to the response are visible, thus, the structures should have a relevance for the neuron behavior. We found in V1-L2/3 mostly Gabor like shapes of a size comparable to the receptive fields of V1-L4 neurons (Fig. 6.9a). Also the neurons in V2 respond to Gabor like stimuli, which in contrast appear larger and less ideally shaped (Fig. 6.9bc). Some neurons developed a selectivity to edges of multiple orientations. The neurons in V2-L2/3 appear sensitive to larger structures than the neurons in V2-L4. Further, they respond to very large edge structures which can be curved. However, the curvature can be an artifact of the lack of large straight contours in the natural scenes in combination with the larger receptive fields of the V2-L2/3 neurons. Also in experimental studies selectivity to obtuse as well as for acute angles was found (Hegdé and Van Essen, 2000, 2003). Although, V2 was described as not specialized to particular contours or textures (Hegdé and Van Essen, 2003). Our results do not disagree with that studies and indicate a selectivity for more complex structures, such as textures or angles, but also a selectivity to simple shapes as bars. Despite this, the observed receptive field shapes can not refine the assumptions about selectivity V2.

Responses to naturalistic textures

The role of V2 neurons remains unclear, since V2 neurons are found to respond on bars, similar to V1 neurons (Freeman et al., 2013). Thus, Freeman et al. (2013) came up with another theory, stating that V2 is sensitive to the dependencies contained in synthetic naturalistic textures, whereas V1 equally responds to these textures and spectrally matched noise images, which do not contain these dependencies (Freeman et al., 2013).

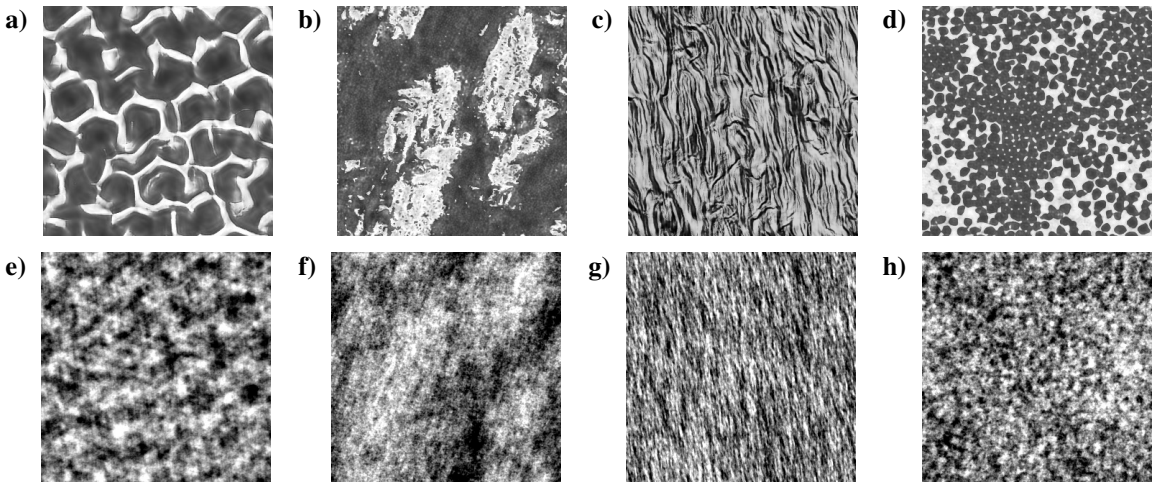

FIGURE 6.10: Examples of synthetic naturalistic textures and the related spectrally matched noise images. Four different synthetic naturalistic textures are shown (top row) and the related naturalistic noise images, having randomized phase values. The images are the same as used in Freeman et al. (2013) and have been provided to us with the courtesy of Corey Ziemba.

Naturalistic textures or visual textures are a sub-class of natural images, which is in general characterized by containing repeated and spatially homogeneous structures with some degree of randomness. Whereas spectrally matched noise images have the same spectrum, but randomized phases (Freeman et al., 2013). Freeman et al. (2013) synthesized 15 different classes of natural textures from different photographs, using the texture synthesis approach of Portilla and Simoncelli (2000). Additionally, they generated spectrally matched noise images by applying a fast Fourier transformation on the original images and replacing the phase values in the spectrum by the phases of uniformly distributed white noise images. The resulting spectrum is transfered back into the image space through inverse Fourier transformation (for examples see Fig. 6.10). In total 450 images have been generated, 225 synthetic naturalistic textures and 225 spectrally matched noise images, consisting of 15 texture families with 15 examples each. The images are in gray scale and have a resolution of 320 by 320 pixel. The original images used in Freeman et al. (2013) have been provided to us with the courtesy of Corey Ziemba.

Our model processes just 24 by 24 pixel image patches, which have been whitened (see Sec. 3.3). Thus, we whitened all images similar to the natural scenes, we are using for network training. This includes an image wise contrast adjustment. Additionally, we equalized the mean contrast of the two groups of synthetic naturalistic textures and spectrally matched noise images. This is because the overall response strength of the model

neurons is contrast dependent and any change in the contrast of one group of images would shift the modulation indexes of all neurons. We made control simulations with different strategies of normalization, but observed qualitatively similar results as the one reported below. We presented 5 by 5 central non overlapping patches of each whitened image to our network, covering 120 by 120 pixel of the original image. We recorded the responses of all model neurons. From the responses, we calculated the modulation index of each neuron for any pair of synthetic naturalistic texture and its spectrally matched noise counterpart. The modulation index is calculated as the difference between the response of a neuron i to a synthetic naturalistic texture r_i^{SNT} and the related spectrally matched noise image r_i^{SMN} and is normalized by the sum of the responses (Eqn. 6.12).

$$MI_i = \frac{r_i^{SNT} - r_i^{SMN}}{r_i^{SNT} + r_i^{SMN}} \tag{6.12}$$

Finally, the modulation indexes of each neuron on the 25 patches, taken from the 225 pairs of images (giving 5625 pairs of patches), are averaged. So that we obtained one modulation index value for each neuron of each network population. From that we calculated the histogram of the modulation indexes of all neurons in each population and compared the distributions to the one of macaque monkeys. We repeated the experiment with 10 independent model runs and report the average modulation indexes and their standard deviations.

Freeman et al. (2013) found that V1 neurons have an average modulation index close to zero, while V2 neurons show a substantial modulation (Fig. 6.11a). We used a more fine grained analyze and measured the modulation index for any neuron population separately. We found that the excitatory neurons in V1 layer-4 as well as in V1 layer-2/3 have a slightly negative average modulation index (Tab. 6.1). The modulation becomes positive in V2, with a minor increase in V2 layer-4 and a substantial increase in V2 layer-2/3 (Fig. 6.11b). Note, the value of the modulation index depends on the contrast of the image families. When we increase the contrast of the naturalistic images, we increase the modulation index of all neuron populations. This changes the measured values, but do not qualitatively change the result that both layers in V1 have a similar average modulation index, V2 layer-4 has a slightly increased modulation index, and V2 layer-2/3 has a substantially increased modulation index.

Additionally to the excitatory neuron populations, which are the most important as they project to higher layers and areas, we analyzed the modulation indexes of the inhibitory

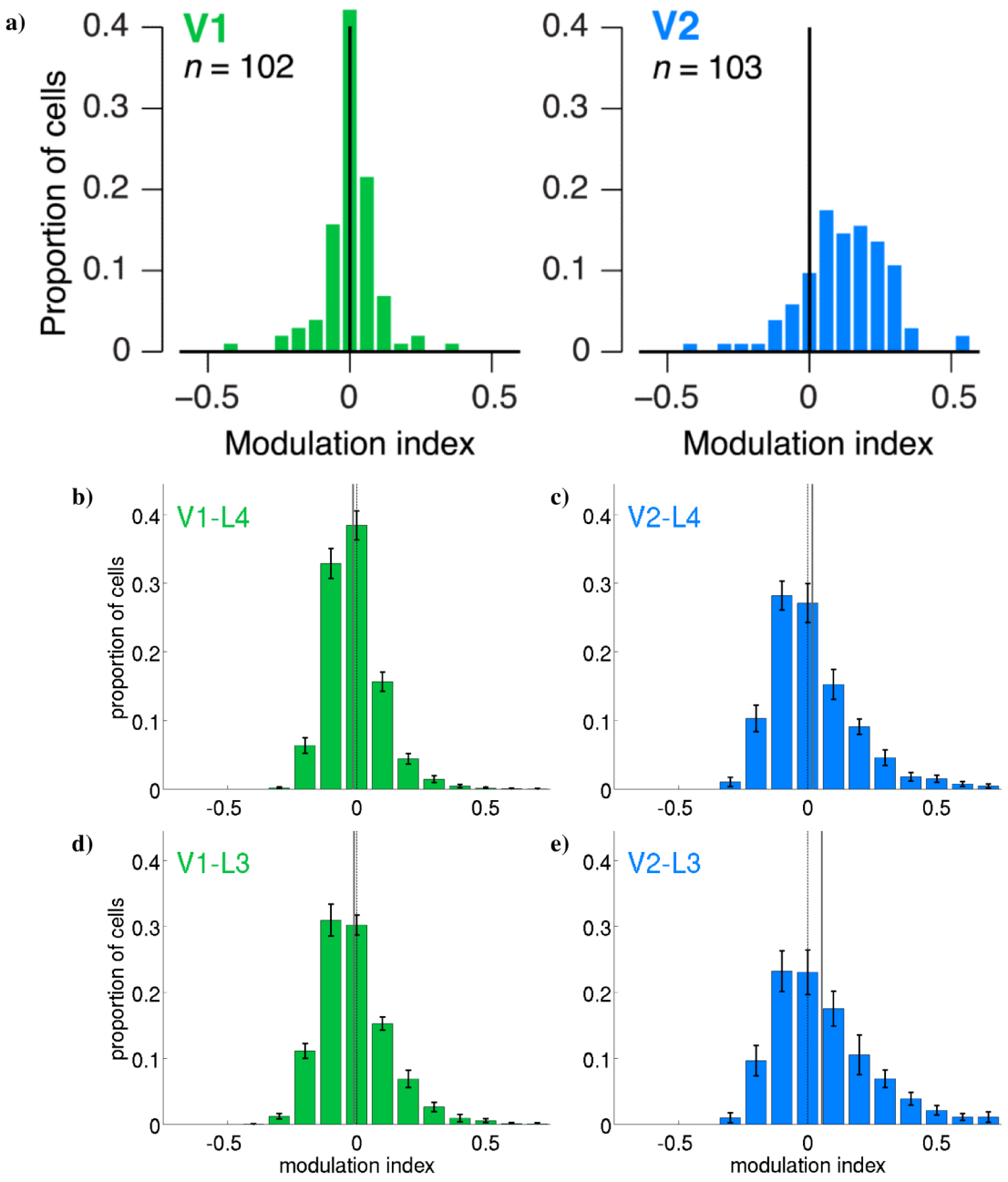

FIGURE 6.11: Histograms of the modulation indexes of macaque monkeys and excitatory network layers. a) Modulation indexes of neurons in macaque V1 and V2, taken from Freeman et al. (2013). **b-e)** Modulation indexes of each excitatory network population, averaged over 10 runs. The error bars indicate ± 1 SD. The gray line indicate the average modulation index. The modulation index of each neuron is calculated as difference between the response onto synthetic naturalistic textures and their spectrally matched noise counterparts, normalized by the sum of both responses and averaged over all images. The data of the macaque show a distribution with an average close to zero, while a majority of neurons in area V2 show a positive modulation. Also both model V1 layers show an average modulation close to zero. First V2 layer-2/3 show a stronger positive modulation. Note, the average modulation of all layers is directly dependent on the contrast normalization of the images. Changes in contrast will shift all model results similarly.

populations. We found that both populations in V1 have similar positive modulation indexes and both in V2 also have similar modulation indexes, which are substantially higher than those measured in V1 (Tab. 6.1). The modulation indexes of the inhibitory populations have been, in general, higher than in the related excitatory populations. This might have its reason in the depth of the inhibitory populations, which is deeper than the one of the excitatory neurons. Thus, the finding that in deeper network layers the modulation indexes are higher might explain the higher values. However, the layer-4 inhibitory populations receive, beside the lateral input from the excitatory neurons, the same feedforward input as the excitatory neurons, which should make their responses more similar to the excitatory neurons. This would also not explain why the layer-2/3 inhibitory neurons show similar values to the layer-4 neurons. There might be a relation to the higher sparseness and lower correlations found in the inhibitory populations (see Sec. 6.3.1). However, it remains unclear why sparser and less correlated neurons should respond stronger to synthetic naturalistic textures (the mean firing rates of the inhibitory neurons are increased by about 10 percent in all layers for naturalistic textures).

Further, we compared the amount of neurons with an average modulation index higher than the average modulation index of the excitatory neurons in V1 layer-4, i.e. we ignore the height of the index which can distort the results when just a minority of neurons showed strong modulations. We found that nearly the same amount of excitatory neurons in V1 layer-4 had modulation indexes above or below the average, just 0.5 percent more neurons had a higher modulation (Tab. 6.1). Similarly, just 2.2 percent more of the excitatory neurons in V1 layer-2/3 showed a higher modulation than the excitatory neurons in V1 layer-4. In V2 layer-4 (excitatory), the number increased to 9.7 percent more and further increases to 28.1 percent more for the excitatory neurons in V2 layer-2/3. This goes in line with the observed increase of the average modulation index in the layers. As expected, more inhibitory neurons in V1 showed a higher modulation than the excitatory neurons in V1 layer-4 (14.7 and 14 percent). The inhibitory neurons in V2 again showed an increased amount of positive modulated neurons (33.4 and 28.3 percent). Interestingly, the amount of inhibitory neurons with high modulations in V2 layer-2/3 (28.3) is quite similar to the amount of excitatory neurons in the same layer (28.1), but the average modulation was found twice as high as in the excitatory neurons (0.12 to 0.06). This is because many inhibitory neurons show stronger modulations in comparison to the excitatory ones. We found also more neurons having stronger negative modulations, but the positive modulated

neurons prevail.

Layer	Type	Mean MI ±SD	rel. inc. MI%
V1-L4	Excitatory	-0.0139 ±0.0027	0.51
	Inhibitory	0.0520 ±0.0039	14.70
V1-L2/3	Excitatory	-0.0100 ±0.0047	2.19
	Inhibitory	0.0505 ±0.0110	14.00
V2-L4	Excitatory	0.0182 ±0.0058	9.71
	Inhibitory	0.1238 ±0.0120	33.38
V2-L2/3	Excitatory	0.0556 ±0.0201	28.13
	Inhibitory	0.1196 ±0.0227	28.33

TABLE 6.1: Average modulation indexes and amount of more positive modulated neurons of each network population. The mean modulation index (MI) ±1 SD is reported. Additionally, the table shows the relative amount of neurons (in percent) which had an increased modulation index in comparison to the average modulation index of the excitatory neurons in V1 layer-4. V2 neurons showed in general a higher modulation than V1 neurons. Inhibitory neurons showed in general a higher modulation than excitatory neurons. A similar relative amount of inhibitory neurons in V2 layer-2/3 had an increased modulation index, but showed stronger modulations.

Similarly to Freeman et al. (2013), we found a higher sensitivity for naturalistic textures in our V2 neurons. Freeman et al. (2013) found this effect in electrophysiological recording in macaque monkeys as well as in human fMRI recordings. Moreover, the effect was present in awake as well as in anesthetized macaques. This indicates that attentional processes play no role and the sensitivity to naturalistic textures is a property of V2 itself. This is important for us as our model can not account for attention related processes. They concluded that this response property robustly differentiates between V1 and V2. Further, they speculated that V2 "complex-cells", which would be attributed to the second stage in V2, namely layer-2/3, better respond to higher-order correlations in the images. Our model results agree with that, as the excitatory V2 layer-2/3 neurons show substantially increased modulations. Moreover, our model would predict that inhibitory neurons show the same effect of increased sensitivity across the hierarchy, but with an in general increased sensitivity for naturalistic textures.

6.3.3 Weight distributions and connection probabilities

Following, we regard the developed connection structure of the network. Since the network consist of 24 different connections (see Chapter 2), we focused on the first and last network

layers. In contrast to our previous work (Wiltschut and Hamker, 2009; Teichmann et al., 2012), all weights in our network are positive defined, the difference between inhibitory and excitatory synapses is implemented in the activation function of the neurons. Thus, we have just positive weight strengths and the distributions appear cut at zero.

We used Hebbian synaptic plasticity rules, which should lead to synaptic weight distributions relative to the linear correlation (covariance) of the connected neurons. Our excitatory, as well as inhibitory learning rule, increases the weights based on the coactivity of the neurons, however, both differ in the normalization part. The second term of the inhibitory learning aligns the weight to the correlation, but with positive weight values for non correlating neurons (see Sec. 6.2.4). In the excitatory learning rule (see Sec. 6.2.3), the weight change is relative to the correlation of the neurons, which can be seen when expanding the first term (the population mean is closely related to the temporal mean of the presynaptic activity, see Sec. 5.3). The Oja normalization term induces an additional constraint on the vector length, which scales the weight strengths relative to the height of the correlation. Thus, the excitatory weights strength should become high for highly correlating neurons and around zero for uncorrelated neurons. In consequence, the weight distributions should reflect the correlation structure of the neurons.

The neurons in the brain have many synapses, but just a few evoke strong postsynaptic potentials (Cossell et al., 2015). Cossell et al. (2015) showed for lateral excitatory connections that this strong synapses develop between highly correlated pairs of neurons. Further, they showed that connections are formed in a non random manner, relative to the degree of correlation of the neurons.

We analyzed the distribution of all of our 24 network connections. We show the weight distributions for the neurons in V1-layer 4 and V2-layer 2/3. Further, we investigated the relation of the developed weight strengths to the response correlation of the connected neurons for the lateral connections between excitatory and inhibitory neurons. Moreover, we show the relation between correlation strength and connection probability, which depends on the interplay between synaptic plasticity and structural plasticity. We compare this data to findings in the primary visual cortex layer-2/3 of mouse (Cossell et al., 2015) and rat (Yoshimura and Callaway, 2005).

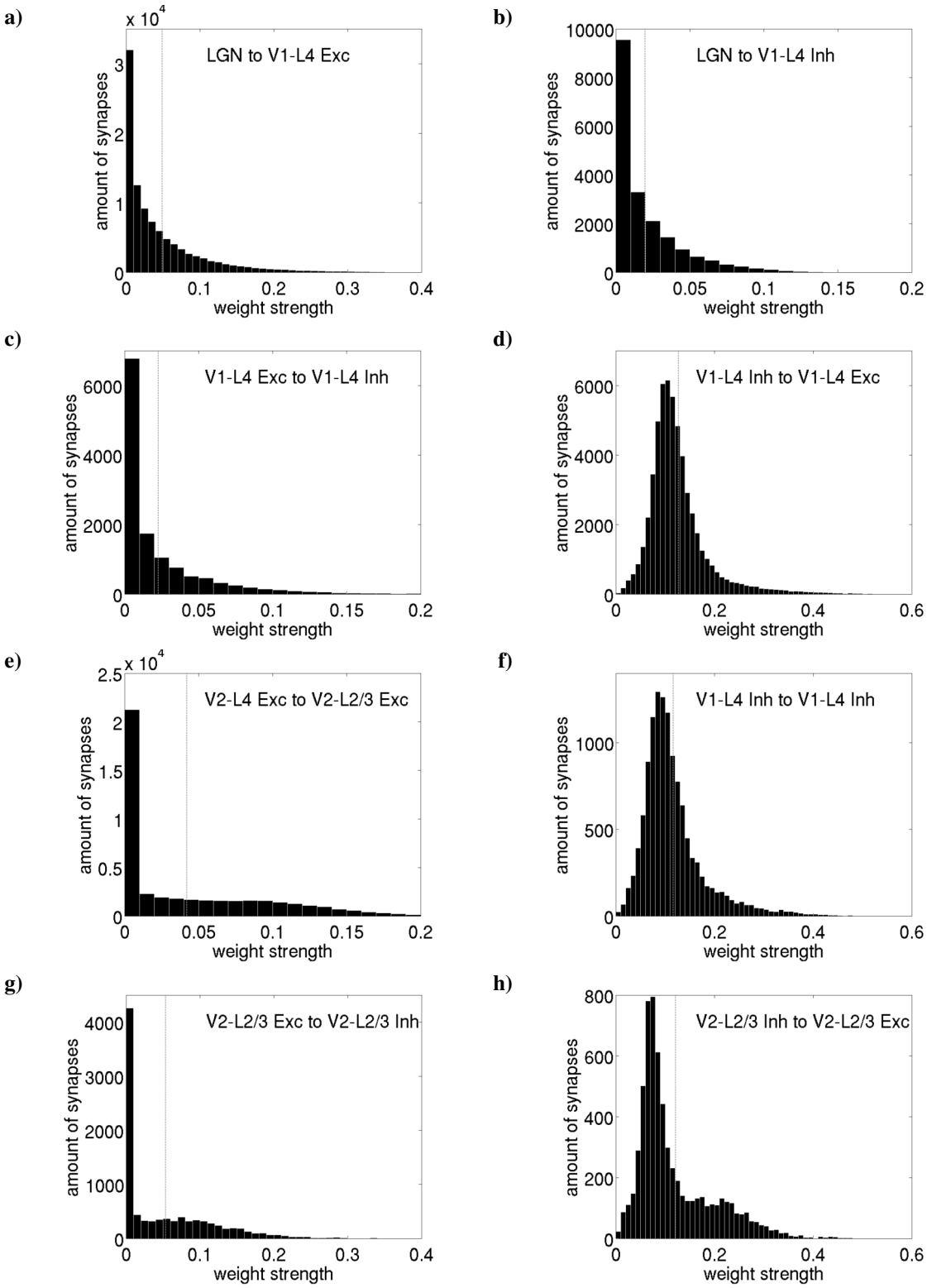

FIGURE 6.12: Weight distributions for selected connections. All histograms have a bin size of 0.01. The dashed gray line indicate the average weight. **a-h)** Selected feedforward and lateral connections. Source and target population are shown in the graphs. When the source population is excitatory the connections are excitatory, analogue for the inhibitory connections.

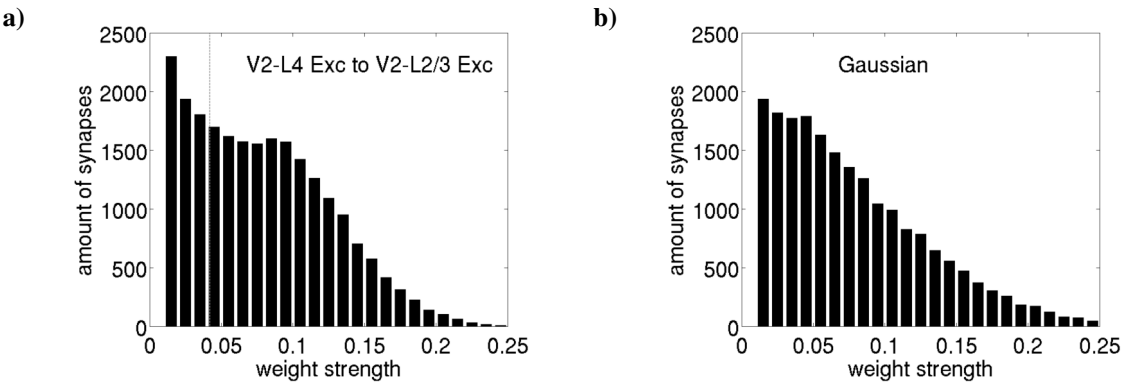

FIGURE 6.13: Comparison of the excitatory weights to layer-2/3 in V2 to a Gaussian distribution. a) Visualization of the weight strength distribution of the excitatory connections from the excitatory neurons of layer-4 in V2 to the excitatory neurons of layer-2/3 in V2. We removed the first bin of the data with values close to zero. **b)** A distribution obtained by randomly selecting the same amount of elements as in a) from a Gaussian distribution. The negative values are mapped to zero and the bin close to zero was not visualized. The mean of both distributions is kept similar.

Weight distributions

The excitatory weight values from the input layer (LGN) to the neurons in V1-L4 appear exponentially distributed (Fig. 6.12ab). With increasing depth this scheme changed gradually to a truncated Gaussian distribution (see Fig. 6.13 for an example) in the deepest connections (Fig. 6.12ceg). Despite that the synapse deletion probability has its peak at zero weights, the excitatory connections showed a large amount of synapses with very low weights. This means the synapse formation counterbalances the synapse removal and each neuron has also in the converged state (for network convergence see Sec. 4.3.5) a large pool of weights allowing him to learn new things. The strength of the excitatory connections is relative to the amount of excitatory connections a neuron receive. For instance the excitatory neurons in V1-layer 4 receive only excitatory input from LGN, whereas the inhibitory neurons in the layer receive excitation over three connections. Hence, the weights of all connections together accumulate to a similar value to the excitatory weights the excitatory neurons in V1-layer 4 receives.

Inhibitory weight values appear log-normally distributed in lower layers (Fig. 6.12df). In deeper layers the log-normal distribution is superimposed by a second distribution with higher mean, this is, probably, the Gaussian distribution observed in the weight structure of the deeper excitatory connections, which explains the right tail of the distribution (Fig. 6.12h). The peak in the weight distribution results from our inhibitory learning rule which

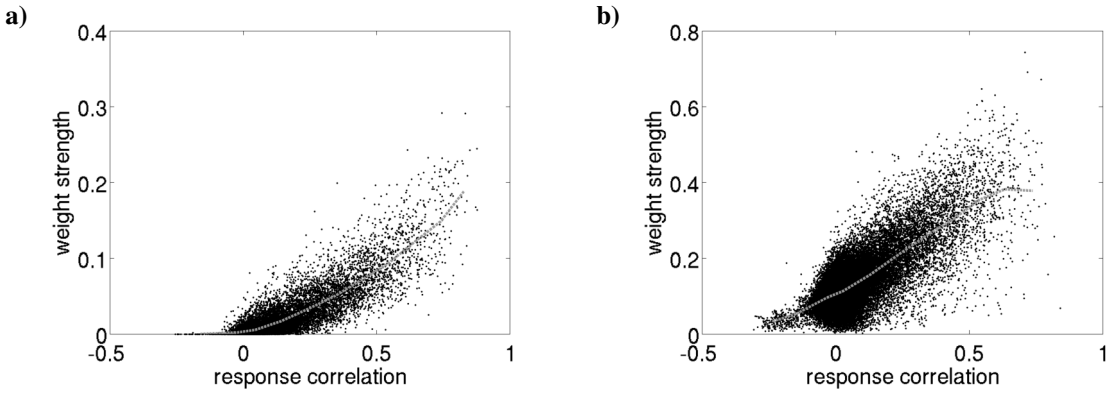

FIGURE 6.14: Weight distributions for the lateral connections of layer-4 in V1 related to the response correlation of the connected neurons. The dashed gray line indicates the average value within a bin of 0.05 with at least 50 elements. **a)** Connections from the excitatory neurons to the inhibitory neurons. **b)** Connections from the inhibitory neurons to the excitatory neurons.

develops positive weights for uncorrelated neurons. The synaptic weights between neurons with zero correlation are in our configuration around 0.1, dependent on the mean activities of the neurons. Additionally, the structural plasticity forms new weights uniformly distributed around the weight value where the removal probability is half its maximum, which is at 0.15 of the normalized weights, thus a bit below 0.1. However, the structural plasticity has just a minor effect on the weight distribution, in model versions without this mechanism the weight distribution appears similar, just the anyway steep decrease of the weight amounts with values close to zero is slightly stronger, because of the synapse removal. Thus, the development of positive weights for the huge amount of uncorrelated neurons in our model leads to the Gaussian like distribution around the weight value of 0.1, whereas stronger weights are depending on the (positive) correlations between the neurons. Which has an exponential tail in lower layers and a Gaussian tail in deeper layers.

Relation between weight strength and response correlation

To get deeper insights in the relation of the weight values to the correlation between the neurons we analyzed the lateral connections between excitatory and inhibitory neurons in all layers. We calculated the response correlations between all neurons (for details see Sec. 6.3.1) and relate it to the weight strength of all formed synapses. For all layers we observed comparable distributions. Excitatory connections developed weight values close to zero for negative response correlations (Fig. 6.14a). The majority of weights are formed

between uncorrelated neurons and have low synaptic weights. With increasing correlation also the weight values increase. As known from the weight distributions, the inhibitory neurons develop mostly positive weight values (Fig. 6.14b). Negatively correlated neurons are again connected by synapses with low weights. As expected from the inhibitory learning rule, weakly correlating neurons developed positive weights with a certain extend of the weight distribution. Between stronger correlating neurons the inhibitory connections are also stronger. In both connection types an increase in response correlation is roughly linear related to the weight value. Similarly to our results a relation between the response correlation and the evoked postsynaptic potential was found for the lateral connections between excitatory neurons in layer-2/3 of the primary visual cortex of mouse (Cossell et al., 2015, Extended Data Figure 2a). Cossell et al. (2015) showed that the evoked average postsynaptic potential is low for weakly correlated neurons and increased with the increase of the response correlation.

Relation between connection probability and response correlation

Until now we analyzed just formed connections in our network. Although, our structural plasticity rule modifies the connection structure based on the learnings of the synapses. It forms new synapses in the neighborhood of existing synapses and removes synapses with low weight values (see Sec. 4.2). Here we link the response correlation of the neurons to the connectivity. Therefore, we calculated the linear correlations of the responses between all pairs of neurons within each network layer (excitatory with inhibitory) on 100000 randomly selected natural scene patches and count the synapses formed between these neuron pairs. For comparison, we report the chance level, which is calculated as the fraction of all formed synapses to all possible synapses (potential pairs). We show the fraction of potential pairs between a synapse can be formed for different levels of correlation. Cossell et al. (2015) found in the primary visual cortex of mouse that the probability of excitatory connections is strongly related to the correlations between the neurons. In contrast, inhibitory connectivity is often reported as unspecific (Harris and Mrsic-Flogel, 2013). Inhibitory interneurons are found to have high connection probabilities to neighboring excitatory neurons, whereas excitatory neurons are found to have much lower probabilities (Yoshimura and Callaway, 2005).

For our lateral excitatory connections from the excitatory to the inhibitory neurons we obtained a similar behavior (cf. Cossell et al., 2015, Fig. 1h). The neuron pairs with

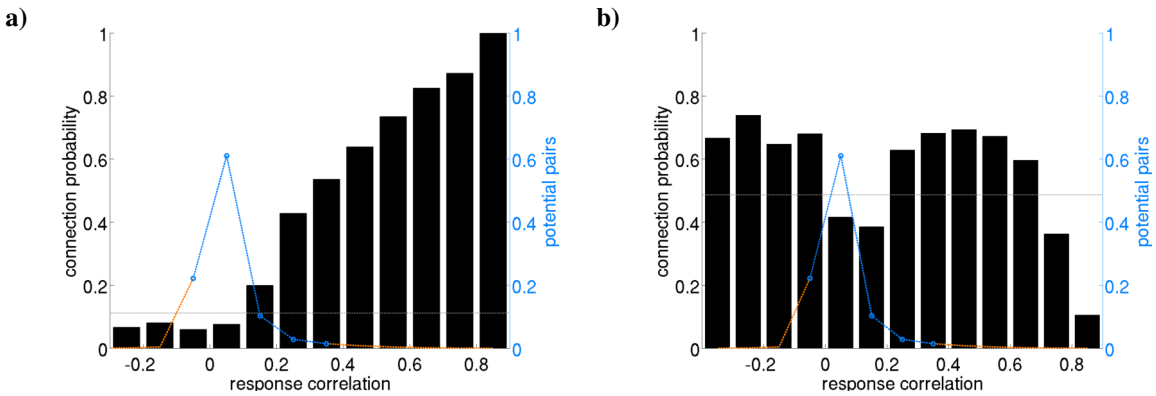

FIGURE 6.15: Connection probability for the lateral connections of layer-4 in V1 related to the response correlations of the neurons. The amount of formed connections between neurons having a correlation value laying in the same bin is related to the amount of neuron pairs with these correlations. Each bin has a size of 0.1. The horizontal gray dashed line indicates the chance level of being connected. The dashed blue and orange line shows for each bin the fraction of potential pairs. For probabilities above 1 percent the line is blue, below it is orange. The most pairs have response correlations close to zero. **a)** Connection probabilities for the connections of the excitatory neurons in V1 layer-4 to the inhibitory neurons in the same layer. The connection probability is below chance level for weakly and negatively correlated neurons and increases gradually with the response correlation. **b)** Connection probabilities from the inhibitory neurons in V1 layer-4 to the excitatory neurons in the same layer. Negatively and positively correlated neurons have connection probabilities above chance level. Weakly correlated neurons have are connected below chance level. The connection probability decreases for highly correlated neurons.

the highest correlations are nearly always connected (Fig. 6.15a). Indeed, just very few neurons show that high correlations (the dashed orange line indicate a fraction below one percent of all possible connections). That is probably the reason that Cossell et al. (2015) reported no neurons with correlations above 0.4. We observed a gradual increase of connection probability for correlations above zero. Uncorrelated or negatively correlated neurons form connections with a probability below chance level (indicated by the horizontal gray dashed line).

The lateral inhibitory connectivity appears odd, in comparison to the excitatory. For negatively correlated neurons we obtained high connection probabilities, as well as for positively correlated neurons (Fig. 6.15b). Weakly correlated neurons, representing the majority of all potential pairs, showed connection probabilities below chance level. Surprisingly, for highly correlated neurons the connection probability dropped far below chance level. The rather high chance level is caused by our limited network size with many neurons with overlapping receptive fields and which cause a need for being inhibitory connected.

The low connection probability for highly correlated neurons can be explained by the special characteristic of inhibition. We have seen that these neurons develop strong weights, however, these weights effect strong inhibition which in turn decreases the correlation. Additionally, these highly correlated pairs of neurons are rare. Thus, when through the structural plasticity a connection is formed to a highly correlating neuron, rapidly strong weights develop, so that the correlation between these neurons decreases shortly after and with that also the weight decreases. As consequence, we have a low probability to observe highly correlated neurons being connected. The one which were connected will be found shortly after in the group of neurons with lower correlations or even negative correlations, until their synapse is removed. But the removal is a slow process, as structural plasticity acts on a much larger timescale than synaptic plasticity. This might contribute to the surprisingly high amount of connections between negatively correlated neurons, with low weight values. Note, substantially negatively correlated neurons are as rare as highly positively correlated neurons. We observed a similar structure of connectivity, as reported above, in all layers. In deeper layers we observed slightly increased connection probabilities, because of the larger receptive field sizes and the increased receptive field overlap.

Our findings for the connection probabilities of the inhibitory connections can explain why many physiological studies report unspecific connectivity (Harris and Mrsic-Flogel, 2013). We found that the connection probabilities are weakly dependent on the correlation between the neurons. We observed in general high values, similar to the one found in physiological studies (Yoshimura and Callaway, 2005). However, we found a strong weight dependency on the response correlations. The aspect that in our model many inhibitory synapses are present between weakly or negatively correlated neurons with positive weight values, can be related to the remarkable amount of untuned inhibition in the visual cortex (e.g. Ringach and Malone, 2007; Xing et al., 2011). It can also be a computational advantage, regarding the highly dynamic process of decorrelation, where a fast buildup of inhibitory weights is important for functioning, i.e. the synapses have to exist beforehand. Moreover, this behavior can explain why modeling studies and some physiological studies found tuned inhibitory neurons (e.g. King et al., 2013; Hirsch et al., 2003), but on the other hand other report an unspecific connectivity to the neighborhood (e.g. Hofer et al., 2011; Harris and Mrsic-Flogel, 2013). We have not analyzed whether some of our neurons form subgroups with specific connections (cf. Hirsch et al., 2003) and others with very broad connectivity as found in physiological studies (cf. Hirsch et al., 2003;

Yoshimura and Callaway, 2005; Hofer et al., 2011; Harris and Mrsic-Flogel, 2013). Nevertheless, we can conclude from the connection probabilities and the distribution of the weight strengths that neurons, which have formed reciprocal connections, will also mostly have strong weights. Something similar is found in rat primary visual cortex, where higher currents between reciprocally connected neurons, than between just one way connected neurons, have been observed (Yoshimura and Callaway, 2005). That is because of the relation of synapse maintenance and synaptic strength to the response correlation. Excitatory neurons have just to highly correlated neurons high connection probabilities and also the connection probability of inhibitory neurons is increased for high correlations. So that together with the higher expectation value for the weight strength of both synapses strong reciprocal connection are likely to originate. Whereas for weakly correlated neurons the connection probability and the expectation value of the weight is low, resulting in a lower average current (cf. Yoshimura and Callaway, 2005, Fig. 1fg, Fig. 3).

From our observations we can further predict the connection probabilities of inhibitory interneurons in the brain. We first predict a decrease in connection probability for highly correlated neurons. Second, we predict high connection probabilities independent from the correlation strength. We link this to the decorrelation through inhibition decreasing the correlations faster than connections are removed. Albeit, our observations are influenced by the formulation of our inhibitory learning rule, which causes a high amount of connections between weakly correlated neurons. With another parametrization, as the one of King et al. (2013), more specificity might be achieved. This has to be tested in further versions of the model.

Contribution of the weights between the most correlated neurons to the total synaptic weight

Another finding was that among the huge amount of synapses the neurons have just few synapses highly contribute to the neurons activity (Cossell et al., 2015). Similar to Cossell et al. (2015), we sorted the weights between excitatory and inhibitory neurons for each network layer by the response correlation the connected neurons have and accumulate the weight values. We normalized the result so that the sum of all weights has a sum of one. For comparison, we accumulated the amount of weights. When all weights would contribute equally both curves would be equivalent.

We found for the lateral excitatory connections to the inhibitory interneurons that less

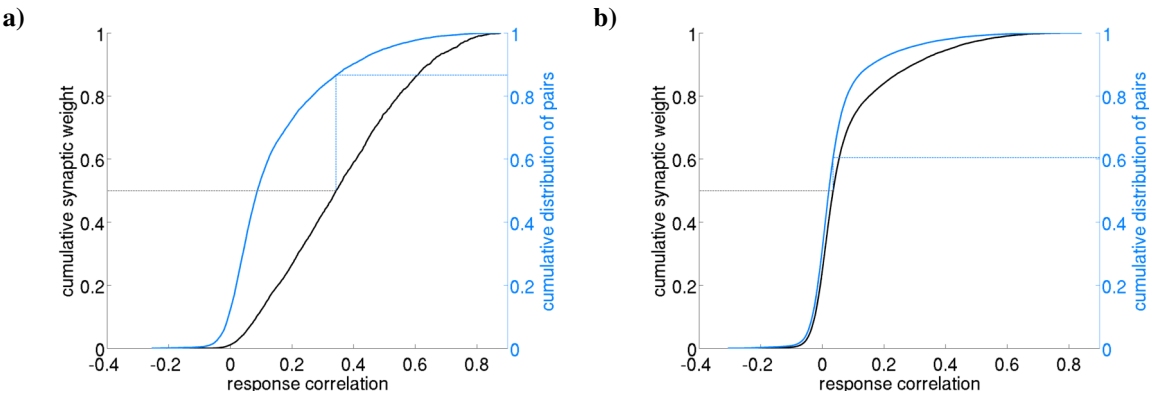

FIGURE 6.16: Weight amount strongly contributing to the total weight. The weight values have been sorted by the response correlations of the connected neuron pairs and cumulated (black curve), indicating the weight contributed by connections between pairs with equal or lower response correlation. The resulting total weight has been normalized to 1. The weight count has been processed similar (blue curve) and indicates how many connections have been formed between pairs of neurons with equal or lower response correlation. The horizontal black dashed line indicates 50 percent of the total weight. The blue dashed line indicates the amount of neurons higher correlations contributing to 50 percent of the total weight. **a)** Connections in V1 layer-4 from the excitatory neurons to the inhibitory neurons. **b)** Connections in V1 layer-4 from the inhibitory neurons to the excitatory neurons.

than 20 percent of the connections (blue curve) contributed to 50 percent of the total weight (black curve) (Fig. 6.16a). This is valid for all network layers. This is related to what Cossell et al. (2015) found in their study, namely that between lateral connected excitatory neurons about seven percent of the connections contributed to 50 percent of the total weight (see Cossell et al., 2015, Fig. 1j). For the connections of inhibitory neurons to excitatory neurons or to the inhibitory population itself we found less specificity. We measured for all regarded connections that 25 to 40 percent of the connections contributed to 50 percent of the total weight (Fig. 6.16b). This again strengthen the idea of the unspecific inhibitory neurons. However, we found a certain degree of specificity, which might change with another parametrization of the learning rule.

6.3.4 Translation invariance

With cortical depth also the invariance of the neuronal responses against the precise position of a stimulus increases (Rolls, 2012). This is called translation (or shift) invariance, because neurons respond similar to translated versions of their preferred patterns. In our model, we address this kind of invariance through the trace learning principle in the layers

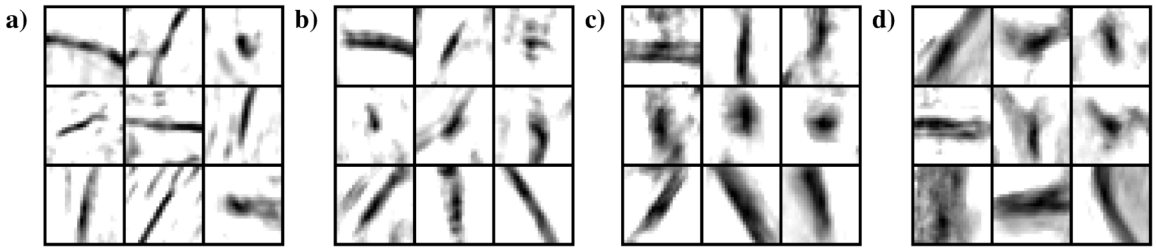

FIGURE 6.17: Example response maps of excitatory neurons from different layers. Each of the nine tiles shows the response map of a neuron from **a)** V1-layer 4, **b)** V1-layer 2/3, **c)** V2-layer 4, **d)** V2-layer 2/3. The responses have been obtained by shifting the optimal natural scene stimulus by 12 pixels in each direction. The gray tone indicates the response strength, where the maximum response of the neuron is black. White indicates no response. Broader tuning indicates increasing translation invariance.

2/3. Thus, we expect increasing invariance in these layers.

To measure translation invariance, we determined the optimal stimulus of each neuron by presenting natural scene patches to the model. We accepted a patch as optimal for a neuron when it evokes a sufficient high response. Subsequently, we shifted the found patch by 12 pixels in all directions, giving us a map of 25 by 25 response values for each neuron. The response maps of V1-layer 4 neurons are intended to be edge like, as their preferred stimuli are edges, whereas in deeper layers the response maps should have a less specific shape, presumably Gaussian shaped with unspecific extents. We fitted the the response maps with a freely rotatable 2D-Gaussian (see Sec. B.3). Because of the elongated receptive field shapes in the early layers the invariance against shifts is expected to be much larger in one dimension than in the other. Thus, we used the smaller of the two Gaussian extents from the fit to characterize the degree of translation invariance. Following we report the results obtained from 10 independent model runs.

We visualized the response maps for the different network populations. We saw mostly sharply tuned edge like response maps for V1-layer 4 neurons (Fig. 6.17a). Whereas the response maps in deeper layers became broader (Fig. 6.17b-d). Note, the structures in the natural scene determine the responses. This means, edges are often larger than the receptive field is and also other structures are present in the image patch. This can lead to response maps with larger structures than the receptive field size. Further, neurons might also respond to structures surrounding the optimal stimulus. Nevertheless, the extents of the main structure gives a good insight in the spatial tuning profile of the neurons.

We compared the distributions of the minimal extents of each network population and

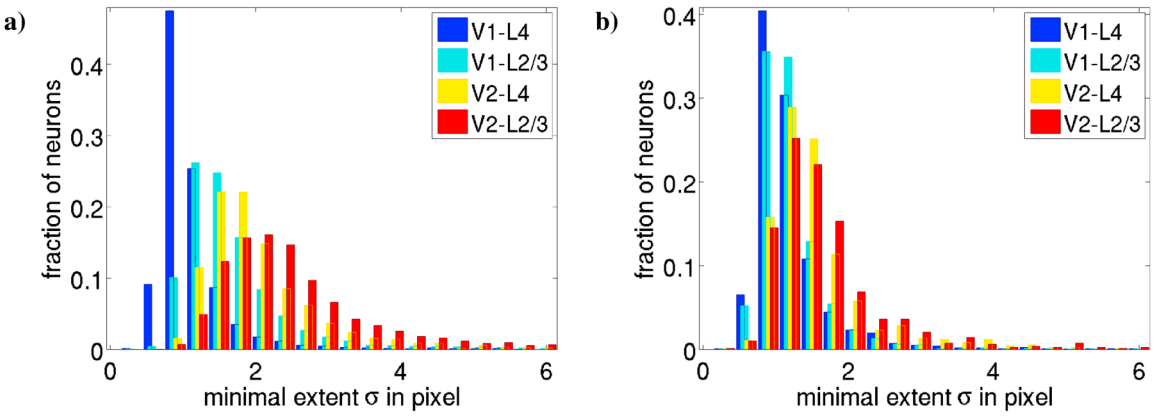

FIGURE 6.18: Distribution of the minimal Gaussian extent for the excitatory neurons of different layers. We fitted a Gaussian to the response maps and used the minimum of the extents as indicator for the broadness of the response map, indicating the degree of translation invariance. **a)** Shows the histogram of the minimal extents for 10 independent runs of the standard model configuration with long trace. **b)** Histogram over three independent runs of a model configuration with short trace. The translation invariance of the model with long trace increases with layer depth. With short trace just a minor increase is observed within an area. V2 showed more invariance than V1 but on a lower level than in the models with long trace.

found that the extents increase with the depth of the network layer (Fig. 6.18a). The median extent of the neurons in V1-layer 4 is around 1 pixel, the extent in V2-layer 2/3 is with more than 2 pixel twice as high (Table 6.2).

In Teichmann et al. (2012) we showed that the trace length in layer-2/3 largely influences the emergence of invariance properties. We compared the response maps of V1-layer 2/3 neurons learned with a long trace with the one learned with a short trace. We found that the long trace led to neurons with more position invariance. In the here presented model we used the same long trance of $\tau_{Ca} = 500ms$ and a similar network training. To gain more insights how far the observed translation invariance is a result of the used trace learning mechanism we repeat the analyzes for a model version with a short trace of $\tau_{Ca} = 10ms$. Note, a time constant for the trace much shorter than the presentation time (100ms) is roughly similar to direct use of the firing rate for learning. Following we report the results of three independent model runs.

We found visually as well as in the distributions of the minimal extents and their median values that the neurons in the model with short trace poorly developed translation invariance (Fig. 6.18b). All populations in V1 are found to have similar distributions of the minimal extents, with similar median values (Table 6.2). Just a minor increase from

Layer	Type	Median of the minimal extents	
		$\tau_{Ca} = 500ms$	$\tau_{Ca} = 10ms$
V1-L4	Excitatory	1.0043	1.0707
	Inhibitory	1.1054	1.0645
V1-L2/3	Excitatory	1.4947	1.1121
	Inhibitory	1.5149	1.1092
V2-L4	Excitatory	1.8404	1.3949
	Inhibitory	1.7628	1.2836
V2-L2/3	Excitatory	2.2587	1.4527
	Inhibitory	2.1420	1.2940

TABLE 6.2: Median values of the minimal Gaussian extents for all populations. We fitted a Gaussian to the response maps and used the minimum of the extents as indicator for the broadness of the response map, indicating the degree of translation invariance. The median values are taken over the data from 10 independent model runs for the long trace $\tau_{Ca} = 500ms$ and three independent model runs for the short trace $\tau_{Ca} = 10ms$. The translation invariance of the model with long trace increases with layer depth. With short trace just a minor increase is observed within an area. V2 showed more invariance than V1 but on a lower level than in the models with long trace.

layer-4 to layer-2/3 was observed, much lower than the one with long trace. Similar to the model with long trace the invariance increased from V1 to V2, but on a lower level. Again in V2, the translation invariance just minorly increased from layer-4 to layer-2/3. Thus, we conclude that the longer trace enables the excitatory neurons in layer-2/3 to develop translation invariance, which in consequence increases the invariance of all subsequent layers. Further, also the increase in receptive field size from V1 to V2 cause higher invariance, however, the increase, with short trace, within an area has just minor effects. The inhibitory neurons behave roughly similar to the excitatory neurons in the same layer. This was expectable from their position in the network hierarchy and the circumstance that we do not use long traces for learning their afferent connections.

6.3.5 Object recognition performance

The visual system, particularly the ventral pathway, enables us to differentiate between different objects. It is hypothesized that with increasing cortical depth the representation, i.e. the neuronal code, of objects becomes better and better linear separable (DiCarlo and Cox, 2007). Accordingly, we evaluated the object recognition performance for each of our

FIGURE 6.19: The 100 objects from COIL-100 in their frontal view.

model populations. We chose the COIL-100 (Fig. 6.19) dataset (Nene et al., 1996)[1], which was often used to determine the invariance against rotations (e.g. Wersing and Koerner, 2003; Einhäuser et al., 2005; Kanan, 2007; Mobahi et al., 2009; Zou et al., 2012). This gives us the additional possibility to see how invariant the representations of deeper layers are. The dataset contains 100 different real world objects photographed in five degree rotations giving 72 poses per object and 7200 images in total. Each image has a resolution of 128 by 128 pixel and RGB color.

We converted all images into gray scale and applied the same whitening procedure as we have used for the natural scenes (see Sec. 3.3). Due to the difference of the image resolution and the input size of the model, we split the images in nine by nine parts with the size of 24 by 24 pixel, using a stride of 13 pixel. We presented all patches to the already trained network and recorded the responses of all neurons. Note, the network was trained on natural scenes and has never seen any image from COIL-100 before. All plasticity mechanisms of the network have been turned off. To obtain sufficient network responses we normalized the contrast of each image so that it evokes, as a whole, a peak response around one. We concatenated the obtained responses of all patches, which belong to one image, to one response vector for each population. So that we obtained a vector of the length 81 times the amount of neurons in the regarded population, representing a single image of the dataset.

[1] `http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php`

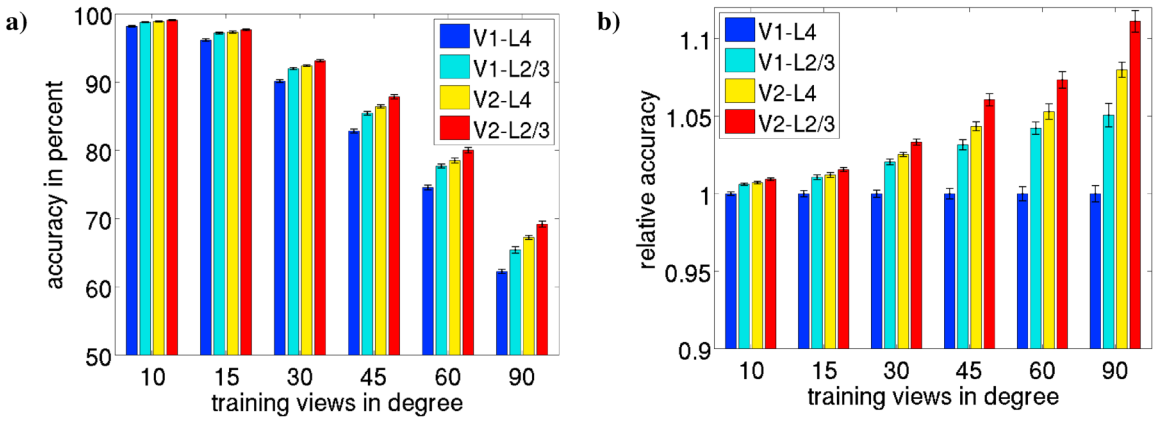

FIGURE 6.20: Comparison of the recognition accuracies on the COIL-100 dataset of different layers. The network has been trained on six different sets of regularly chosen views and tested on the unseen views. For instance every second view, which means that the training views cover each 10 degree of the object rotations. **a)** The recognition accuracies for the excitatory neurons in all network layers in percent. **b)** The relative gain in accuracy to the excitatory neurons of the first network layer V1-L4.

For training the classifier, we split the response data into a training and test set. We used six different splits: every 2nd view (10°), every 3rd (15°), 6th (30°), 9th (45°), 12th (60°), and every 18th (90°). Which gives us 36, 24, 12, 8, and 4 training views per image. The test set contained all other views. With that we trained a support vector machine (SVM) using a linear kernel and we predicted the labels of the test images and calculated the recognition accuracies. As software we used *libsvm*[2] in the version 3.21. For comparison we applied the same evaluation method on the raw color images, the gray scaled images, and the whitened gray scaled images (for all accuracies see App. D.2).

In the easiest case, with training on every second image view, we achieved 96.4 percent accuracies on the network input. The excitatory neurons in the first network layer V1-layer 4 achieved a higher accuracy of 98.2 percent (averaged over 10 independent model runs) and in the deepest network layer 99.1 percent (Fig. 6.20a). The accuracies increased monotonically with layer depth under all training conditions. With increasing difficulty (training of the classifier on less views) the overall recognition accuracies decreased. Interestingly, under more difficult training conditions, requiring more rotation invariance, the excitatory neurons in the deeper layers showed an increased gain in accuracy over the first layer neurons (Fig. 6.20b). This might have its reason in the increased invariance of these

[2] `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`

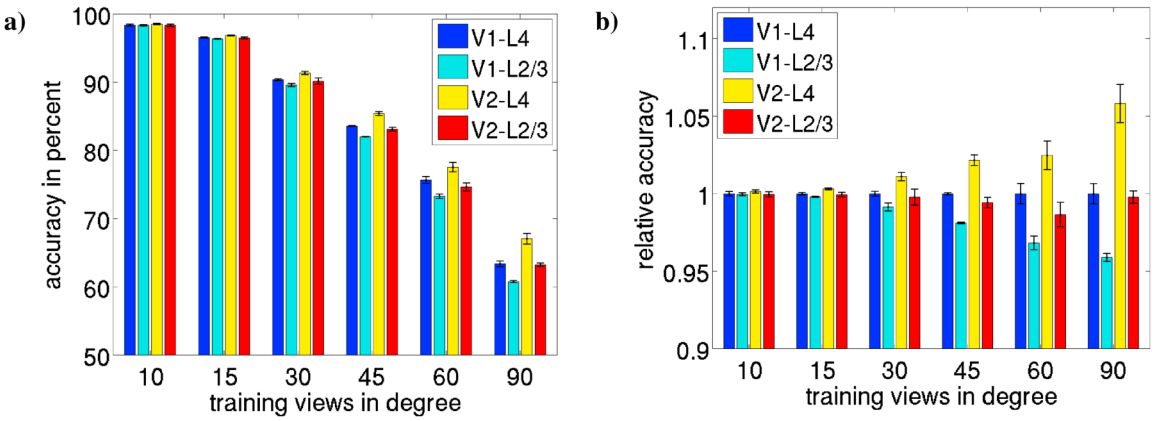

FIGURE 6.21: Comparison of the recognition accuracies on the COIL-100 dataset of different layers. The network has been trained on six different sets of regularly chosen views and tested on the unseen views. For instance every second view, which means that the training views cover each 10 degree of the object rotations. **a)** The recognition accuracies for the excitatory neurons in all network layers in percent. **b)** The relative gain in accuracy to the excitatory neurons of the first network layer V1-L4.

neurons (see previous section). Note, the neurons have not been trained on rotated images, but rotation also has translation parts.

The inhibitory neurons showed a worser recognition accuracy than the excitatory neurons in the same layer (Table 6.3). This is presumably caused by the lower amount of neurons of about one fourth. However, inhibitory neurons are not projecting to higher areas, thus, their accuracy is of minor importance. We also observed no improvement when using the combined neuronal code of inhibitory and excitatory neurons. A detailed overview on the recognition accuracies of all network populations and layers, including the standard deviation of the accuracies obtained in the 10 independent model runs, can be found in Appendix Table D.2.

The achieve recognition results are in range with other approaches using this dataset. Kanan (2007) reported on gray scaled images for 10 training views a recognition accuracy of 76.26 percent and for 15 views 78.87 percent. Therefore, they combined Gabor filters with eccentricity dependent size with a visual memory and a simple Bayesian classifier. We obtained with 8 training views (45°) between 82.8 (V1-L4) and 87.9 percent (V2-L2/3) recognition accuracy. Mobahi et al. (2009) used a deep learning approach to generate an invariant feature set. Therefore they used temporal coherence as supervisory signal. When they trained their deep network on just the labeled training set, i.e. the colored images of the COIL-100 dataset, they obtained with 4 training views (similar to our 90° case) a

recognition accuracy of 71.49 percent, slightly better than what we obtained on gray scaled images (69.2 percent). However, with using the rotations in the dataset as continuous input stream to learn an invariant feature set they obtained 92.25 percent, much higher than we do. Note, in this case the test images have been part of the training set, thus, the temporal coherence learning will learn features well matching to all dataset images. (Zou et al., 2012), also a deep network, learned invariant features by a temporal slowness constraint. They achieved with that features recognition accuracies on gray scaled images of 87 percent (4 training views). Wersing and Koerner (2003) obtained 79 percent (4 views) on the colored images. They employ a deep network with feature, learned with sparse invariant feature decomposition, and pooling layers converging to view tuned units. Wersing and Koerner (2003) as well as Mobahi et al. (2009) listed earlier results of other authors for a SVM classifier (polynomial kernel) with 74.6 percent and a nearest neighbor classifier with 70.1 percent (4 training views), worser than what we obtained with a SVM and linear kernel on colored images (85.5 percent, see App. Table D.1).

To gain deeper insights on the observed relative increase in recognition accuracies in deeper layers of our network for more difficult conditions, i.e. the classifier is trained on less views, we again compared our network with the version with short trace. The network with short trace did not show the same monotonic behavior of increasing recognition accuracies with deeper layers (Fig. 6.21). Particularly, when the classifier was trained on few views the excitatory neurons in the layers 2/3 showed a drop in performance in comparison with their preceding layer. Whereas the first layer V1-layer 4 already showed a comparable accuracy of 98.4 percent (averaged over three independent runs), the maximum accuracy remained by 67.1 percent, two percent off in comparison with the standard model (Table 6.4).

Thus, we concluded that the more invariant features (see previous section) are beneficial to the object recognition in deeper layers. With the use of a long trace in the layers 2/3, we obtained a substantial increase in translation invariance and an improvement in recognition accuracy of rotated objects, which indicates an improved rotation invariance. The neuronal representations learned with short trace in the layers 2/3 are disadvantageous for object recognition, but contained enough information on the objects to learn better representations in subsequent layers. The strong increase of object recognition accuracy in V2-layer 4 is presumably caused by the larger receptive fields of these neurons, having higher translation invariance (see previous section). The neurons in V2-layer 2/3 showed

Layer	Type	Training views					
		10°	**15°**	**30°**	**45°**	**60°**	**90°**
Network input		96.4444	94.1250	87.0333	78.3906	71.1970	58.7206
V1-L4	Excitatory	98.2000	96.1833	90.1550	82.8469	74.5682	62.2676
	Inhibitory	97.9083	95.7604	88.2433	80.4375	71.9909	60.2559
V1-L2/3	Excitatory	98.8028	97.2083	92.0033	85.4609	77.7182	65.4235
	Inhibitory	98.7778	97.0896	91.6667	84.9484	76.8333	64.3397
V2-L4	Excitatory	98.9028	97.3479	92.4367	86.4453	78.5182	67.2441
	Inhibitory	98.7000	96.9646	91.2900	84.5641	76.5030	64.7588
V2-L2/3	Excitatory	99.1139	97.6833	93.1567	87.8641	80.0439	69.1926
	Inhibitory	99.0083	97.4333	92.3500	86.6297	79.1803	67.9162

TABLE 6.3: Recognition accuracies on COIL-100 of all network populations. We trained a SVM with linear kernel on the images from the COIL-100 dataset, using particular views. We tested on all unused views. We show the mean accuracies for each network population, obtained from 10 independent network runs.

just a minor increase in translation invariance, together with the lower amount of neuron the learned representations are disadvantageous over the preceding layer. We further saw that the linear separability of the object representations becomes easier with increasing depth of the layers, when we used a long trace. We also observed this effect when we use a short trace, but with stronger fluctuations between the layers.

6.4 Conclusion

We proposed a set of synaptic plasticity mechanisms within a framework able to learn biologically plausible neuron and connection properties. Further, we addressed the main functional feature of the visual cortex, invariant object recognition. For the excitatory synapses in our network we used a combination of covariance learning and Oja normalization, together with a postsynaptic calcium trace. The calcium trace was configured with different trace lengths, a short for the layers 4 and a long for the layers 2/3. The short trace acts similar to the direct use of the firing rate. The length of the long trace is selected so that it averages over the responses on multiple stimuli. It is intended to exploit the temporal continuity of the visual input to form invariant representations. The Hebbian learning principle was completed by anti-Hebbian learning of the inhibitory synapses. The inhibition should serve the required amount of differentiation in the neuronal activity of correlating neurons, which in turn differentiates the patterns learned by the excitatory synapses and serves decorrelated neuron responses. Together with the intrinsic plasticity, which aims

Layer	Type	Training views					
		10°	**15°**	**30°**	**45°**	**60°**	**90°**
V1-L4	Excitatory	98.3796	96.5556	90.3667	83.5885	75.6768	63.3775
	Inhibitory	97.8148	95.6250	87.7667	79.9271	71.3889	59.3333
V1-L2/3	Excitatory	98.3426	96.3611	89.5944	82.0104	73.2778	60.7745
	Inhibitory	97.9537	95.7153	88.0833	79.8125	70.7273	58.8333
V2-L4	Excitatory	98.5185	96.8542	91.3667	85.3958	77.5505	67.0637
	Inhibitory	97.8056	95.6458	87.9056	80.2135	71.9697	60.6029
V2-L2/3	Excitatory	98.3426	96.5000	90.1667	83.1094	74.6566	63.2402
	Inhibitory	96.5556	93.5417	84.7222	76.4479	68.1010	57.3627

TABLE 6.4: Recognition accuracies on COIL-100 of all network populations for the model with *short trace*. We trained a SVM with linear kernel on the images from the COIL-100 dataset, using particular views. We tested on all unused views. We show the mean accuracies for each network population, obtained from three independent network runs. The model is similar configured to our standard model but uses a short calcium trace ($\tau_{Ca} = 10$), so that the trace not effects learning.

equal mean and variance of the activities, the inhibitory plasticity allows the neurons to represent the complete input space. Inhibitory learning is self stabilizing, because the neurons go over in a winner-take-all competition, where the inhibitory weight is not increasing anymore. Excitatory plasticity can increase the weights as long as the postsynaptic firing rate becomes not too high. However, inhibition decreases the firing rate. To avoid a race between both plasticity forms, we limited the length of the weight vector by controlling the α value of the Oja normalization, based on the excitation a neuron receives. Further, we allowed the neurons to determine their desired individual weight vector length by themselves. We constantly increase the allowed weight vector length and use a squared term to reduce the length, when the firing rates breach an upper threshold. With that, we implemented compensatory processes on multiple time scales, one instantaneous, in form of the Oja normalization, and one on a much slower time scale, modulating the overall weight vector length. For the need of such a rapid and a slow compensatory process was argued by Zenke and Gerstner (2017). They based their argumentation on several experimental results which found slow compensatory processes, whereas mathematical models employ fast compensatory processes to achieve stabilized Hebbian learning.

We found that with that combination of Hebbian and anti-Hebbian learning an efficient neuronal code can be formed. We achieved high population sparsenesses as well as high lifetime sparsenesses of the individual neurons. This is accompanied with low correlations between the neurons within a population. We argued that the achieved code is not a result

of an explicit goal carried out by the elements of the neuronal system. Rather it is the natural consequence of the combination of Hebbian excitatory and inhibitory plasticity, where inhibitory plasticity decorrelates the neurons which causes sparseness.

We compared the receptive fields learned by our network neurons to physiological findings and computational models. We could show that our model V1-layer 4 neurons developed Gabor-like filters as well as blob-like filters. Early ICA and sparse coding models developed Gabor-like filters, but no blob-like filters, which have been found in monkeys. We showed that our inhibitory V1-layer 4 neurons developed similar receptive fields as the excitatory one. To gain insights in deeper network layers, we used the back projection of the weights into the image space. We showed that V1-layer 2/3 neurons developed connections to V1-layer 4 neurons with similar orientation tuning.

For deeper layers we advanced this technique by weighting the connection matrices with the responses evoked by an optimal natural scene patch. We could show that V1-layer 2/3 neurons respond for edge-like structures with a limited extent. Also V2 neurons tend to respond for edge-like structures, but also for more complicated texture-like parts. The size of the structures, to which the neurons have been sensitive, increased with layer depth. The regarded properties are in range with findings of physiological studies which also found V2 selective to edges and more complex structures as angles or textures.

We also tested a newer theory which claims that V2 neurons, in contrast to V1, are more sensitive to naturalistic textures than to spectrally matched noise, which has the same spatial frequencies but differs in the arrangement. We found, as suggested, no specific tuning for naturalistic textures in V1. But V2 neurons showed an increased average modulation to naturalistic textures. Further, we made a more fine grained comparison, where we found a smaller increase in the average modulation for V2-layer 4 and a stronger increase in V2-layer 2/3. Surprisingly, the inhibitory neurons in all layers had an higher modulation indexes than the corresponding excitatory neurons. The largest difference was found in V2-layer 4. The observed differences in average modulation where also accompanied with an increased amount of neurons showing higher modulations, compared to the average of the excitatory neurons in V1-layer 4. This suggests differences between cell type and cortical layer and should be considered in future experimental studies.

We further analyzed the weight structures in our network. We found that the feedforward excitatory weights in the first layer followed an exponential distribution. Whereas, the distribution of the (non zero) excitatory weights in deeper layers became more and

more similar to the top half of a Gaussian distribution. This is presumably caused by the change in the correlation structure. In early layers the activity correlations are strongly influenced by the correlations of the network input, whereas in deeper layers the correlations are shaped by the interplay of excitation and inhibition. The distribution of the inhibitory weights differs from the excitatory ones. In early layers it appeared log-normally distributed. In deeper layers they appear also log-normally distributed, but imposed by a Gaussian distribution.

Subsequently, we analyzed the relation of the weight strength to the neuron response correlations. We found that our excitatory plasticity rule developed weak to zero weights for uncorrelated or negatively correlated neuron pairs. With increase of the correlation also the weight value increased. Inhibitory plasticity showed a similar behavior, with increasing correlation the weights increased. However, it developed higher weight values for uncorrelated neurons. This correlation dependent weight strength could be linked to experimental data, showing that the evoked postsynaptic potential for weakly correlating neurons is low and increases with the degree of correlation of the neuron pairs.

Since our synaptic plasticity interacts with the structural plasticity, we further analyzed how the correlation structure, determining the weight strength, is related to the connectivity structure. Experimental data suggested that with increasing correlation between the neurons also the connection probability increases. We could confirm this in our model for the excitatory connections. We found connection probabilities below chance level for uncorrelated or negatively correlated neurons. With increasing correlation also the rate of connection increased. However, highly correlating neuron pairs ($corr > 0.4$) are rare. For inhibitory connections the picture was more complex. Inhibition directly influences the correlation of the connected neurons. We observed a higher connection probability for negatively correlated neurons than the chance level. Uncorrelated neurons and weakly correlated neurons had a connection probability below chance level. Medium high correlated neurons again had a connection probability above chance level. Surprisingly, we found highly correlated neurons connected largely below chance level. This odd appearing distribution was pointed to two effects. First, a similar effect as for the excitatory neurons, which let the connection probability increase for higher correlations. This is quite logical as we applied the same structural plasticity algorithm on both connection types. The second effect is the genuine functioning of the inhibition. As seen before, the inhibitory weights tend to be positive also for low and negative correlations. Thus, when neurons at one point

in the development had a high correlation, they will most likely form a connection, when the connection is formed the inhibition reduces the correlation, as consequence they might end uncorrelated or negatively correlated. Because of the much slower timescale of the structural plasticity these connections remain for a long time. The other effect is that unconnected neurons can develop a highly correlated firing before a connection is formed. When a connection is formed the correlation will quickly decrease. Thus, despite highly correlating neurons should have high connection probabilities, just few connected neurons can be observed. The consequence is that the connections of inhibitory neurons appear unspecific and connected to everything around, as it was reported in several experimental studies, despite they underly the same specificity as the excitatory one. We found no qualitative differences between layers or areas.

It appears logical that with a long tailed distribution of the weight strengths, as an exponential distribution, few strong weights largely contribute to the total weight a neuron receives. We also observed that strong weights where formed between highly correlating neurons. On the other hand, the majority of the potential synapses has been found between uncorrelated or weakly correlated neurons and had just a low connection probability. In a physiological study, it was shown that the weights between the few most correlated neurons contributed the half of the total weight. We also could show for the excitatory neurons in our model that the synapses between the few most correlated neurons provided the half total weight. Also inhibitory connections showed this property, but here up to 40 percent of the most correlated connections, dependent on the layer, where needed to provide 50 percent of the total weight. We linked the remaining large amount of weakly or uncorrelated correlated neurons with positive weight value, contributing the other 50 percent of the total weight, to the experimental finding of a large fraction of so called untuned inhibition.

To gain deeper insights in the invariance properties of our network we evaluated the translation invariance of each neuron. Note, the network was trained on a continuous stream of natural scene patches with small random translations. We found a consistent increase in translation invariance with increasing layer depth, where inhibitory neurons behave similar to the corresponding excitatory neurons. However, our trace learning mechanism, which enforces invariance learning, just effects the layers 2/3. Thus, we compared our network model to a version with short trace, ineffective to learn from temporal coherences. We found that the layers 2/3 showed no relevant gain in translation invariance anymore. However, V2-layer 4 already had an increase in invariance. We related this to

the increase in receptive field size of these neurons.

We have the aspiration that the model fulfills the primary function of the visual system, object recognition. Further, we followed the hypothesis that deeper layers form a neuronal code where objects are better linear separable. Consequently, we measured the recognition accuracy of all network populations separately. Therefore, we presented images from the COIL-100 dataset to our network, which was trained on natural scenes, and trained a linear SVM with the network responses. We observed an increasing recognition accuracy in deeper layers. Further, when we trained the classifier on less object views, the other views could be better classified with the responses of the deeper layers. This indicates that the neurons in the deeper layers have, beside their increased translation invariance, also an increased rotation invariance. We compared the results again with the network with short trace and found that the layers 2/3 drop in their performance, in comparison to their preceding layers, while V2-layer 4 improved its performance, which was presumably caused by the already observed increased translation invariance of this layer.

7 General Discussion

In this thesis we combined three important plasticity mechanisms within one model of the early visual system. With synaptic plasticity we implemented a biological plausible form of self-organization. We treated homeostatic mechanisms, effecting the synapses, as part of the synaptic plasticity. We demonstrated the importance to combine synaptic plasticity with intrinsic plasticity, which stabilizes the operating point of the neurons and enables distributed informative representations. Finally, we enhanced the model with structural plasticity, which allows us to examine the relation between the learnings and the connectivity. It also has the advantage to reduce the modeler's bias on the results and helps to form suitable connection matrices between the manifold connected neurons.

7.1 Intrinsic plasticity

Intrinsic plasticity is the key element of our model implementation. Controlling the neuron's operating point to have equal mean and variance enabled us to implement a multilayer network with Hebbian plasticity. We could show that controlling the operating point, improves the information coding in deeper layers. With that, the learnings are not dominated by a few strongly active neurons. However, intrinsic plasticity was originally introduced as computational mechanism for maximizing the neuron's information (Stemmler and Koch, 1999; Triesch, 2005a,b), which can be achieved through an exponential distribution of the neuron activities, given a fixed mean in the case of a single neuron (Simoncelli and Olshausen, 2001). Particularly in the work of Triesch (2005a), and later, an exponential response distribution was used as objective to modify the neurons nonlinear activation function to transfer its activities to an exponential output distribution. Similar to these approaches we also used a modifiable activation function to allow regulations of the neuron's excitability. But differently, we used a rectified linear activation function, which is not able to transfer any input distribution into an exponential distribution. Moreover, our goal was not to obtain an exponential activity distribution. It was enabling all network neurons

to participate equally in the encoding of the stimuli. Hence, our implementation has no objective to enforce an exponential distribution of the activities. Indeed, we also achieved an exponential distribution. But we could show that our intrinsic plasticity mechanism is not responsible for that. Without using intrinsic plasticity the firing rates follow an exponential distribution. We attribute the emergence of an exponential response distribution to the inhibitory plasticity in our network. Inhibition is modeled to reduce the correlations between the neurons, which in turn reduces linear dependencies and with that enforces independence. In contrast to a single neuron setup, information is maximized in a multi neuron setup when all neurons are independent (Simoncelli and Olshausen, 2001). Which not contradicts exponential response distributions. With decorrelation also comes sparseness, where just a few neurons are active at the same time. Thus, neurons have rarely high activities, as high values are rare in an exponential distribution. This makes an exponential distribution a natural outcome of the circuit functioning. Such distribution was also found in a model for V1 simple-cell learning in response to natural scenes (Falconbridge et al., 2006). The model combines, similar to us, Hebbian learning with anti-Hebbian learning of the lateral inhibition, without having intrinsic plasticity mechanisms.

We argue that for a successful implementation of a multi neuron setup two ingredients are important: decorrelated activities and neurons equally participating in the stimulus encoding. Intrinsic plasticity serves just the latter. Moreover, we argue that equal participation is required for deep networks with Hebbian plasticity rules local to the synapse, because similar response statistics of the presynaptic neurons are required to avoid biases. Without that a few neurons would dominate the postsynaptic activity and as consequence the encoded information.

Intrinsic plasticity has been originally demonstrated for individual neurons (e.g. Stemmler and Koch, 1999; Triesch, 2005a,b, 2007; Joshi and Triesch, 2009). Later it was successfully used it in multi neuron setups (e.g. Butko and Triesch, 2007; Savin et al., 2010; Neumann et al., 2013; Diehl and Cook, 2015). However, the improved information encoding in deep networks has not been shown in other models yet.

We have not compared our implementation of intrinsic plasticity to physiological data. Obviously our mechanism changes the sensitivity and the activation threshold, as it was found in experimental studies (e.g. Desai et al., 1999). However, it was reported that intrinsic plasticity takes place over hours up to days (Turrigiano et al., 1998; Desai et al., 1999; Zhang and Linden, 2003; Turrigiano and Nelson, 2004; Nelson and Turrigiano, 2008; Tur-

rigiano, 2011). Whereas our mechanism operates over a few thousand stimulus presentations, which is in the range of several minutes. Further experiments have to show if a slower form could also be effective. Even so, it could be the case that processes on multiple time scales are required, similar to what is hypothesized for synaptic homeostasis (Zenke et al., 2017; Zenke and Gerstner, 2017). In our model Oja normalization would account for rapid homeostatic processes, intrinsic plasticity for mid-term processes, and the regulation of the allowed weight vector length is a long-term process (which means in our model it takes hours). We found it important that intrinsic plasticity acts on a similar timescale as synaptic plasticity, because it aims to reduce imbalances between the neuron activities, which in turn influence the learnings. The synaptic homeostatic mechanisms are local to the neuron and can not reduce such imbalances.

7.2 Structural plasticity

We implemented structural plasticity as stochastic process of synapse formation and removal. The probability for synapse formation is determined by the strengths of the neighboring synapses. The probability is high when many strong synapses are in the surrounding, and low with few or weak synapses around. The probability for synapse removal depends only on the synapse strength. Low weights have high removal probabilities and strong weights negligible low probabilities. Structural plasticity should facilitate learning without the restrictions of the initial connectivity, defined by the modeler. With that it should overcome the modeler's bias on the results. We demonstrated that we can obtain similar learnings outgoing from different starting conditions. We render this property as important aspect to model deep networks with a rich connectivity. That is because neuroscientific data are often just available for early layers, single neuron and connection types such as feedforward connections to excitatory neurons, and can differ between the used measurement methods. Thus, designing a functioning connectivity, which does not bias the learning results, is very difficult. Another important aspect for our model was the preservation of the retinotop organization. We showed that randomly creating synapses in the neighborhood does not harm the retinotop organization. The most neurons developed receptive fields close to their initial connection matrix, which we had retinotop organized.

To our knowledge no other model implemented a related form of structural plasticity in a model of the early visual system. Despite that, many other models share similari-

ties to the mechanisms employed in our model. Several models relate, similar to us, the process of synapse removal to the synaptic weight (e.g. Zheng et al., 2013; Fauth et al., 2015b; Deger et al., 2018). However, some remove the synapses immediately when the weight becomes zero (Zheng et al., 2013; Deger et al., 2018), which appears odd when considering that spine pruning is a slow process (Sala and Segal, 2014). Contrarily, others treat the removal as pure stochastic process (Yasumatsu et al., 2008; Butz et al., 2009a; Fauth et al., 2015a). For synapse formation, some models also consider the neighborhood to existing synapses (Butz et al., 2009a; Butz and van Ooyen, 2013; Butz et al., 2014a,b; Fauth et al., 2015b). Other models use again a pure stochastic process, regardless of the location of the newly formed synapses (Deger et al., 2012; Zheng et al., 2013; Deger et al., 2018). Which appears not suitable for large scale models of the visual system, as the probability that new synapses are formed close to existing ones is low, but the receptive fields are spatially compact. Further, the retinotop organization could be easily impaired when creating several synapses far away from the initial ones. An aspect we have not considered for our model is the firing rate homeostasis through structural plasticity (Butz and van Ooyen, 2013; Butz et al., 2014a,b; Gallinaro and Rotter, 2018). In contrast to our model these models use constant synaptic weights, thus, no other homeostasis can take place. Our model has several mechanisms to regulate the postsynaptic firing rate, namely intrinsic plasticity and synaptic scaling. However, our model can be easily extended to account for homeostatic aspects of structural plasticity. The maximum probability of the synapse removal could be increased when a neuron has too strong activities and decreased when it is weakly active. Similarly, the scaling constant of the formation probability could be increased for weakly active neurons and decreased for highly active neurons.

Despite we use several well supported assumptions for the design of our model, it remains an open question whether structural plasticity is a pure stochastic process or there are some guiding factors. It has been shown that spine formation can be induced by LTP in the dendritic arbor, however, it remains unclear if this is also accompanied by the formation of new synapses (Sala and Segal, 2014). We argue that nothing more than random outgrowth and formation in vicinity of existing synapses, which in turn is related to an active dedritic arbor, and the logical process that weak synapses disappear is required. In our model, we form new synapses dependent on the synapse strength in the neighborhood. The synapse strength is a direct result of LTP and LTD. Thus, we found it effective when potentiation causes synapse formation. The formation of new synapses, however, does not guarantee

their maturation. After formation, we subject them to synaptic plasticity and apply further structural changes only after a certain time interval. Just synapses who sufficiently increase their weight can remain stable. We have shown that synapse strengths at the border of the receptive field are often weak and also when these synapses are likely to be formed they will have high removal probabilities. This aspect might explain the experimental finding of LTP triggered formation and the low degree of persistence and maturation (Sala and Segal, 2014). However, synapse formation requires not necessarily on the weight strengths of the surrounding synapses. The removal of weak synapses, together with the spatially compact shape of receptive fields, would lead to fewer synapses at locations where neighboring synapses are weak. Hence, the synapse amount in the neighborhood alone could indicate the formation probability in a similar way. Our assumptions would predict that dendritic arbors with few synapse would have a low rate of synapse formation.

Besides this, it was found that the change of dendritic arbors is very small in adult neurons (Holtmaat and Svoboda, 2009). Our results seem to confirm this. Once the neurons have learned a proper receptive field, the structural changes become minimal. This is because strong synapses are unlikely to be removed and certainly these synapses determine the postsynaptic activity and with that they determine the learning of the neuron. Consequently, new distant synapses are unlikely to gather sufficient strong weights and are likely to be removed, thus, the dendritic tree can hardly expand. This is visible in our results of the stability of the weight vectors after an initial period of strong plasticity. It was also reported that a brief postnatal phase of synapse formation, with increasing synapse densities, is followed by a period of synapse pruning (Holtmaat and Svoboda, 2009). Also in our model we measured in the early phase of learning, with strong weight changes, a brief increase in the synapse amount, followed by a period of synapse pruning. Indeed, we can not exclude that the definition of our initial weight values causes the brief increase of synapses, but we observed this time course for the most of our connections.

Conclusively, structural changes are linked to the receptive fields of the neurons. This might explain why different cell types and brain regions are found to have different change rates (Holtmaat and Svoboda, 2009). Further, we argue that no guidance is required to explain experimental findings, besides the vicinity of an axon to a dendritic arbor and the removal of weak synapses.

7.3 Synaptic plasticity

We proposed a set of Hebbian plasticity rules. The excitatory learning combines covariance learning with Oja normalization and a postsynaptic calcium trace. We used long trace in the second stage of an area. In the first stage, we used a short trace, similar to a rule without trace. This is related to the well known concept of alternating stages of simple- and complex-layers (e.g. Fukushima, 1980; Riesenhuber and Poggio, 1999; Masquelier and Thorpe, 2007; Kheradpisheh et al., 2016). To enable the excitatory neurons to learn different patterns, we implemented inhibitory interneurons and used anti-Hebbian learning for their inhibitory connections. Both learning rules developed weights relative to the correlation of the neurons. In contrast to previous models, all connections in our network are plastic and learned in parallel. Moreover, deep networks with learned complex-layers are rare. The most deeper models apply a max-pooling stage to obtain translation or scale invariance (e.g. Riesenhuber and Poggio, 1999; Serre et al., 2007; Masquelier and Thorpe, 2007; Kheradpisheh et al., 2016, 2018). However, this requires an organization of the features in the previous layers. This organization can be created by using fixed weights (Riesenhuber and Poggio, 1999), or a plasticity which learns the same features at different positions, for instance the imprint learning of HMAX (Serre et al., 2007) or the learning of one representational feature for any position, as in networks using convolutions (Fukushima, 1980; Masquelier and Thorpe, 2007; Kheradpisheh et al., 2016, 2018). The only multilayer network which also learns invariance over its hierarchy, known to the author, is VisNet (e.g. Wallis and Rolls, 1997; Rolls and Milward, 2000). This networks applies, as we do, trace learning to obtain invariant representations. Also similar to us, it has a retinotop organization of its connections and four layers. In contrast to us, it uses the same learning rule for all layers and makes no differentiation between simple- and complex-layer. However, its first layer has a predefined connectivity, so that the network starts with a “complex-layer”. Albeit, the model has been criticized for its poor performance when learning from natural images (Masquelier and Thorpe, 2007). Further, they proposed that with continuous transformation learning VisNet can also learn invariances without trace learning (Stringer et al., 2006). Despite all the obtained results with the different versions of VisNet (Rolls, 2012), the performance on realistic datasets has to be doubt. The model was neither trained on natural scene images, as our model, and no biological plausible receptive field have been reported, nor the model has shown its capabilities for object recognition on established object recognition datasets. We see the advances of our model in comparison to VisNet

in the more realistic inhibition and the intrinsic regulations, which should lead to a better input representations. The excitatory learning is difficult to compare, as we rely on similar principles. However, we use a much more complex learning rule and could demonstrate, in this work but also with previous models (Wiltschut and Hamker, 2009; Teichmann et al., 2012; Kermani Kolankeh et al., 2015), that we learn proper receptive fields, comparable to the one of primates.

In general there have been few deep neural networks for simulating the visual cortex published, which use biologically inspired unsupervised learning rules. The network of Masquelier and Thorpe (2007) uses a very simple STDP learning rule, but has just one plastic layer. Similarly, the newer version of the network (Kheradpisheh et al., 2016). First with the most recent version, all, now three, convolutional layers are learned (Kheradpisheh et al., 2018). With this version a comparison of the recognition performance to state of the art DNNs was possible, with convincing results. It was also reported that the first layer learns receptive fields, which appear roughly like simple-cells. However, no model version has been trained on natural scene images and when training on object recognition databases than deeper layers show no comparable results to experimental data. Also the complex-layers are non plastic max-pooling layers. Thus, the model can give just very limited insights in the processing of the visual cortex.

An interesting approach from the deep learning community was an unsupervised learned two layer sparse deep belied network (Lee and Ng, 2008). Its two layers are trained on natural scene images. The first layer forms expectedly simple-cell receptive fields. The second layer was described to mimic V2 receptive fields, in terms of contours, corners, and junctions. Beyond the non biological learning scheme, the two layer architecture does not match the layout of the assumed cortical circuit, where V2 is reached after at least two stages of processing. However, they compared their second layer receptive fields to the experimental data of Ito and Komatsu (2004) and achieved some good results. Which might indicate that a hierarchy, as we implement, is not necessary to explain V2 data. Similarly, Spratling (2012) required just two stages of processing to obtain V2 like receptive field properties. However, both approaches differ largely in their employed form of plasticity from classical local Hebbian plasticity, where our learning rules are derived from.

Beside the few deep networks, several shallow network on the level of V1-L4, which utilize more convincing synaptic plasticity mechanisms, have been proposed (Savin et al., 2010; Zylberberg et al., 2011; Willmore et al., 2012; King et al., 2013; Miconi et al.,

2016). All of these networks could demonstrate the learning of simple-cell receptive fields and some also simulated inhibitory interneurons (King et al., 2013; Miconi et al., 2016). However, it remains unclear if their learning principles are applicable in a multilayer network, remind that we required intrinsic plasticity for deeper models. Also none of them showed that invariance properties can be learned with their rules.

In comparison to other models, we presented in this thesis the only deeper model of the areas V1 and V2, which has a certain degree of biological plausibility and demonstrates the learning of receptive field properties comparable to the properties found in primate V1 and V2, including invariance properties of V1 complex-cells. Moreover, we could show the object recognition capability of the model, despite we have been limited in this work by the small input size of our network. Furthermore, we implemented a much more complex neocortical like connection structure, which should be extended in subsequent work to a full recurrent network with cortical feedback (see Teichmann and Hamker, 2017).

Further, we addressed in our evaluations several questions of coding and organization. The knowledge on the model principles allowed us to explain the potential mechanisms behind several neuroscienetific findings. We analyzed basic properties of efficient coding in our model, namely sparseness and correlations. We found high sparseness values and low correlations. Our obtained sparseness values are lower than in other computational models (Hoyer, 2004; Wiltschut and Hamker, 2009; King et al., 2013) and difficult to compare to experimental data. The results vary largely between experiments, presumably because of different levels of anesthesia or animal age (Berkes et al., 2009), or non-sensory inputs to the neurons (Harris and Mrsic-Flogel, 2013). However, in general are the measured sparseness levels similar (Berkes et al., 2009; Willmore et al., 2011) or lower (Berkes et al., 2009) than ours, contrary to the other models. We addressed the different character of the two kinds of sparseness, lifetime sparseness and population sparseness. For efficient coding of visual patterns the population sparseness is relevant, because it measures the amount of responses, i.e. energy, the encoding requires. However, experimentally lifetime sparseness of single neurons is measured and can not in general be interchanged with population sparseness (Lehky et al., 2005; Berkes et al., 2009; Willmore et al., 2011; Spanne and Jörntell, 2015). In our model lifetime and population sparseness behave quite similar. We attributed this to the intrinsic plasticity mechanism which equalizes the neurons basic response characteristics, mean and variance, and enforces an equal participation in the encoding of stimuli. Lifetime and population sparseness can be similar when the neuron

responses are independent (Berkes et al., 2009; Willmore et al., 2011). Thus, we conclude that our neuron responses achieved a certain degree of independence. When we assume similar physiological mechanisms to our model mechanisms, we would predict that also in experimental studies, with mature healthy animals, lifetime sparseness and population sparseness are interchangeable. The assumed certain degree of independence from the similarity of lifetime and population sparseness of our model neurons is also supported by the measured low linear correlation values. The level of decorrelation which we reach is similar to the level found in computational studies with related plasticity algorithms (Wiltschut and Hamker, 2009; Zylberberg et al., 2011; King et al., 2013). In experimental studies the level of correlations again differs, but was also found to be low (Ecker et al., 2010; Smith and Kohn, 2008). We attribute the obtained high sparseness values and low correlations to our inhibitory mechanism. Inhibitory learning does actively reduce correlations between the neurons by increasing their inhibitory weight relative to the correlation. Further, we have seen that neuron populations with lower correlations have sparser responses. Thus, inhibitory plasticity serves a process of active sparsification and decorrelation.

Beside statistics on the neuronal code, we also analyzed the evolved weights of the different neuron layers. We showed that the first layer neurons developed simple-cell like receptive fields. Therefore, we reconstructed the receptive fields from the neuron responses via reverse correlation or used the feedforward weights as good estimate of the receptive fields. We found that our synaptic plasticity, together with structural plasticity, leads to receptive fields of various orientations and sizes. Interestingly, the receptive field size and shape can largely differ from small blob like receptive fields to elongated Gabor-like receptive fields. We showed that the receptive field distribution of our model is more realistic (in comparison to macaque monkey data) than the receptive fields obtained with early sparse coding approaches (Olshausen and Field, 1996, 1997) or models using independent component analysis (Bell and Sejnowski, 1997; van Hateren and van der Schaaf, 1998; van Hateren and Ruderman, 1998). Further, we showed that also V1-L4 inhibitory neurons developed simple-cell like receptive fields. However, the reverse correlation reveals a certain degree of sensitivity to more complex structures, which is presumably caused by their richer connectivity, compared to the one of excitatory neurons. We went beyond the analysis of the V1-L4 receptive fields and visualized the receptive fields of the excitatory neurons in the deeper layers. We used the feedforward weights of the neurons for visualization. For V1-L2/3 neurons, we found that that the neurons learned to connect

predominantly to V1-L4 neurons with a similar oriented receptive field, but a slightly different receptive field position. However, some neurons also developed weights to neurons with different orientations. This indicates that the visual hierarchy might be less strict than assumed and selectivity to more complex features can evolve in earlier or later stages. The receptive fields of our model V2 neurons seem to support this assumption. We back projected the weight matrices and weight it with the network responses on the optimal stimulus of the neurons. This highlighted the structures in the input which evoked strong responses in the neurons. Similar to experimental data (e.g. Hegdé and Van Essen, 2000, 2003) we found a large fraction of neurons responding best to Gabor-like stimuli, as V1 does. However, our data also showed an increase in the complexity which the optimal stimuli can have. Future work has to investigate in more detail the variability and complexity of the visual patterns for which our model neurons are sensitive. In this thesis we investigated another novel hypothesis for the V2 sensitivity. It was proposed that V2 is sensitive to the dependencies contained in naturalistic textures, whereas V1 responds equally to any image composed with the same spatial frequency statistic (Freeman et al., 2013). We also found this effect with our model, although weaker. V1 neurons showed a slightly negative preference for naturalistic textures and V2 neurons showed a positive preference for naturalistic textures. Interestingly, we found a stronger preference in V2-L2/3. Unfortunately it was not reported from which layer in V2 the experimental data have been recorded.

We also regarded the learned weights between the different neuron populations. We found exponentially distributed weights for excitatory connections, which is comparable to the distribution of evoked postsynaptic potentials in lateral excitatory connections (Cossell et al., 2015, Fig. 1d). However, in deeper model layers the distribution gradually changed. We found that the inhibitory weights follow a log-normal distribution. Also here the distribution changed slightly in deeper layers. We do not found that structural plasticity largely influenced the weight distribution, it just reduced the amount of weak synapses by a small fraction. We attributed the change in the weight distribution to a change of the correlation structure in deeper layers, which is presumably caused by our inhibitory learning. The responses of early layers strongly depend on the input distribution, but in deeper layers the inhibitory learning shapes the response correlations.

We already know from the definition of our learning rules that the weight values should be related to the correlation of the activities of the connected neurons. We related the weight strength to the response correlations and could prove this relation for both synaptic

plasticity rules. The learning of the excitatory neurons differed from the inhibitory learning, as it led to zero weights for a large fraction of uncorrelated or negatively correlated neurons, whereas the inhibitory learning led to higher weight values for low or negative correlations. This is caused by the definition of our inhibitory learning rule. A version which should lead to zero weights for uncorrelated neurons has been proposed by King et al. (2013). However, for inhibitory connections positive weights between uncorrelated neurons might be beneficial. Decorrelation is a highly dynamic process which requires a similar time course as for excitatory learning, so that inhibitory weights can be build up fast enough. This is likely to require preexisting connections. Further, it can be the case that the untuned inhibition from these preexisting inhibitory weights cause these uncorrelated activities. Interestingly, untuned inhibition, where we add inhibition from uncorrelated neurons, was found to largely contribute to orientation selectivity (Xing et al., 2011).

We advanced the relation of the weights to the response correlations by considering the relation of the response correlations to the existence of the synapses. We know from our network mechanisms that synapses are removed when they are weak. However, we found for inhibitory synapses that more synapses than by chance are formed to negatively correlated neurons. We explained this by the mentioned highly dynamic process of inhibitory plasticity, which leads quickly after connection to weakly or negatively correlated neurons. Thus, when we regard pairs of neurons with negative correlations, their synapses are likely to be removed, but structural plasticity is a slow process so that we can observe them for a long time period. Thus, we found more of them than expected. Contrarily, highly correlated neurons had a low probability of being connected. We explained this by the opposite effect. Which is that they just become highly correlated when they where not connected and, contrarily to what we would assume from the strong weights which developed for highly correlated neurons, connections are rare. Because when they are connected they will quickly reduce their correlation. Thus, they are just for a short time period observable. This might explain why the connectivity of inhibitory interneurons is reported as less specific (Hofer et al., 2011; Harris and Mrsic-Flogel, 2013), despite inhibitory learning rules should lead to specific weights and tuned receptive fields, which are also found in experimental studies (e.g. Hirsch et al., 2003). We applied the same analysis on the excitatory connections. We found that they form highly specific connections. Negatively and uncorrelated neurons (the majority) have connections probabilities below chance level. The connection probability increased to 100 percent for highly correlated neurons. This

also implies that strong excitatory connections are very stable. Our results for the excitatory connections are very similar to the lateral excitatory connections observed in mouse V1-L2/3 (Cossell et al., 2015). Fortunately, we can measure all connections in our system. Hence, we could also state connection probabilities for the rare connections between highly correlated neurons and confirm the trend seen in the experimental data. Further, we could apply this analysis on all network connections. We found comparable results for all excitatory connections and all inhibitory connections. Which indicates that neurons and layers having the same plasticity mechanisms should show the same connection statistic.

Subsequently, we analyzed the contribution of the weights between the strongly correlated neurons to their total weight. We could confirm the finding that the few connections from highly correlated neurons contribute the most to the total weight (Cossell et al., 2015). For inhibitory neurons we found that also weights between weaker correlated neurons contribute much to the total weight. Which again underpins why these connections are often described as unspecific and inhibition was found to have a large untuned component.

Finally, we addressed the most remarkable property of the visual system. The ability for invariant object recognition. First, we measured the robustness (invariance) of the neuron responses to shifts of their optimal stimuli, the translation invariance. We found translation invariance gradually increasing with depth, similar to the findings with VisNet (Wallis and Rolls, 1997) and following the common hypothesis about the visual system that invariance is build up gradually (DiCarlo et al., 2012; Földiák, 1998). Interestingly, we found that when we did not use long traces in the layers 2/3, that these layers showed a minor increase in translation invariance. We still measured an increase in invariance from area to area, but V2 reached just the level of V1-layer 2/3 with long trace. This indicates that a cortical mechanism, as longer calcium traces, in the second stage of processing within an area would be beneficial to achieve more invariant representations. Subsequently, we measured the object recognition accuracy on the COIL-100 dataset under different difficult conditions. We achieved very high recognition accuracies for the easiest conditions. Remember, our network was not trained on these stimuli, contrarily to other models (e.g. Fukushima, 1980; Kheradpisheh et al., 2016, 2018). Thus, the features learned by our model have to be universal enough to encode the COIL objects. That V1 implements something like a general codebook to encode visual scenes is widely accepted. Because V2 is not understood very well, this aspect is also unclear. Models which are used for object recognition seems to quickly form more object related features, as single views (e.g.

Kheradpisheh et al., 2018). As the V2 features learned on natural scenes give good results on the COIL-100 dataset, we conclude that this indicates that V2 has learned a general codebook, suitable for a large variety of visual stimuli. Under more difficult test conditions, we observed an increasing advantage of deeper layers in recognizing unseen views. These views can be largely rotated in comparison to the training data of the classifier. This further indicates that the features in deeper layers are also more invariant to rotations, besides their increased invariance to translations. Moreover, it proves the capability of our learning algorithms to learn useful encodings of visual scenes, including that also deeper layer contain sufficient information, and invariance learning, i.e. trace learning does not harm information coding. We could further show that without trace learning the recognition performance in the layers 2/3 is impaired. Which renders our applied learning scheme again as beneficial for object recognition. This contrary to the assumption of continuous transformation learning that minor transformations in the input space could lead to invariance learning without the need of a trace (Stringer et al., 2006). We used small random shifts in our training protocol and despite this the layers 2/3 could not build up a comparable degree of invariance. Interestingly, V2-L4 showed in all model versions improved invariance and recognition accuracy over its previous layer.

7.4 Achievements

We could obtain V1 and V2 receptive fields, comparable to assumptions about the receptive field properties in the related visual cortex areas. In contrast to many other models, we also learned the invariance properties of the layers 2/3 and we have shown that trace learning is beneficial for this purpose. To our knowledge, we are the only multilayer model of the visual system, which combines synaptic plasticity with intrinsic and structural plasticity. We could demonstrate that structural plasticity is beneficial in overcoming the bias of the initial connection structure for the learnings. Further, we found that the induced stochastic changes of the structural plasticity do not impair the stability of the network. We showed the positive impact of intrinsic plasticity on the development of the neuronal code in deeper layers and argued that equal participation of the neurons is the important criteria to maximize the represented information of the population, in contrast to forcing an exponential response distribution within individual neurons. Further, we found that we need no objective for achieving an efficient sparse code, inhibitory plasticity, which

reduces the correlations, is sufficient enough. We have not tuned the parameters of our model to fit experimental data. We chose them with the purpose to obtain stable learning and following general assumptions about the relation of the parameters to each other, e.g. the learning speeds, the speeds of the homeostatic regulations, and the probabilities of synapse formation. Despite this we obtained, with the combination of synaptic and structural plasticity, a comparable connectivity statistic to a recent neuroscientific study. Moreover, we could explain why inhibitory connections appear unspecific, although they develop specific weights. Finally, we demonstrated the invariant object recognition capability of our model and showed that invariance is build up gradually, which is impaired without trace learning. Which renders our network design of consecutive layers with either fast trace or slow trace advantageous.

Bibliography

Abbott, L. F. and Nelson, S. B. (2000). Synaptic plasticity: taming the beast. *Nat. Neurosci.*, 3:1178–83.

Adelson, E. H. and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.*, 2(2):284–99.

Albrecht, D. G. and Hamilton, D. B. (1982). Striate cortex of monkey and cat: Contrast response function. *J. Neurophysiol.*, 48(l).

Anderson, J. C. and Martin, K. A. C. (2009). The Synaptic Connections between Cortical Areas V1 and V2 in Macaque Monkey. *J. Neurosci.*, 29(36):11283–11293.

Angelucci, A., Levitt, J. B., Walton, E. J. S., Hupe, J.-M., Bullier, J., and Lund, J. S. (2002). Circuits for local and global signal integration in primary visual cortex. *J. Neurosci.*, 22(19):8633–46.

Anzai, A., Peng, X., and Van Essen, D. C. (2007). Neurons in monkey visual area V2 encode combinations of orientations. *Nat. Neurosci.*, 10(10):1313–21.

Baddeley, R. J., Abbott, L. F., Booth, M. C. a., Sengpiel, F., Freeman, T., Wakeman, E. a., and Rolls, E. T. (1997). Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proc. R. Soc. London. Ser. B Biol. Sci.*, 264(1389):1775–1783.

Barlow, H. B. (1961). Possible Principles Underlying the Transformations of Sensory Messages. In Rosenblith, W. A., editor, *Sens. Commun.*, volume 1, chapter 13, pages 217–234. MIT Press Cambridge, MA, USA.

Barlow, H. B. (2001). Redundancy reduction revisited. *Network*, 12(3):241–53.

Bell, A. J. and Sejnowski, T. J. (1997). The ”independent components” of natural scenes are edge filters. *Vision Res.*, 37(23):3327–38.

Berkes, P., White, B., and Fiser, J. (2009). No evidence for active sparsification in the visual cortex. *Neural Inf. Process. Syst.*, 22:108–116.

Berkes, P. and Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *J. Vis.*, 5(6):579–602.

Bi, G. Q. and Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.*, 18(24):10464–10472.

Bienenstock, E. L., Cooper, L. N., and Munro, P. W. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosci.*, 2(1):32–48.

Binder, M. D., Hirokawa, N., and Windhorst, U., editors (2009). *Encyclopedia of Neuroscience*. Springer-Verlag Berlin Heidelberg, volume 1 edition.

Brette, R. (2015). Philosophy of the spike: rate-based vs. spike-based theories of the brain. *Front. Syst. Neurosci.*, 9(151):1–14.

Butko, N. J. and Triesch, J. (2007). Learning sensory representations with intrinsic plasticity. *Neurocomputing*, 70(7-9):1130–1138.

Butz, M., Steenbuck, I. D., and van Ooyen, A. (2014a). Homeostatic structural plasticity can account for topology changes following deafferentation and focal stroke. *Front. Neuroanat.*, 8(October):1–20.

Butz, M., Steenbuck, I. D., and Van Ooyen, A. (2014b). Homeostatic structural plasticity increases the efficiency of small-world networks. *Front. Synaptic Neurosci.*, 6(April):1–14.

Butz, M. and van Ooyen, A. (2013). A Simple Rule for Dendritic Spine and Axonal Bouton Formation Can Account for Cortical Reorganization after Focal Retinal Lesions. *PLoS Comput. Biol.*, 9(10):e1003259.

Butz, M., van Ooyen, A., and Wörgötter, F. (2009a). A model for cortical rewiring following deafferentation and focal stroke. *Front. Comput. Neurosci.*, 3(August):10.

Butz, M., Wörgötter, F., and van Ooyen, A. (2009b). Activity-dependent structural plasticity. *Brain Res. Rev.*, 60(2):287–305.

Caporale, N. and Dan, Y. (2008). Spike timing-dependent plasticity: a Hebbian learning rule. *Annu. Rev. Neurosci.*, 31:25–46.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., and Rust, N. C. (2005). Do we know what the early visual system does? *J. Neurosci.*, 25(46):10577–97.

Caroni, P., Donato, F., and Muller, D. (2012). Structural plasticity upon learning: regulation and functions. *Nat. Rev. Neurosci.*, 13(7):478–90.

Castellani, G. C., Quinlan, E. M., Bersani, F., Cooper, L. N., and Shouval, H. Z. (2005). A model of bidirectional synaptic plasticity: from signaling network to channel conductance. *Learn. Mem.*, 12(4):423–32.

Cho, K., Aggleton, J., Brown, M., and Bashir, Z. (2001). An experimental test of the role of postsynaptic calcium levels in determining synaptic strength using perirhinal cortex of rat. *J. Physiol.*, 532(2):459–466.

Clopath, C., Büsing, L., Vasilaki, E., and Gerstner, W. (2010). Connectivity reflects coding: a model of voltage-based STDP with homeostasis. *Nat. Neurosci.*, 13(3):344–52.

Cormier, R. J., Greenwood, a. C., and Connor, J. A. (2001). Bidirectional synaptic plasticity correlated with the magnitude of dendritic calcium transients above a threshold. *J. Neurophysiol.*, 85(1):399–406.

Cossell, L., Iacaruso, M. F., Muir, D. R., Houlton, R., Sader, E. N., Ko, H., Hofer, S. B., and Mrsic-Flogel, T. D. (2015). Functional organization of excitatory synaptic strength in primary visual cortex. *Nature*, 518(7539):399–403.

Cummings, J. A., Mulkey, R. M., Nicoll, R. A., and Malenka, R. C. (1996). Ca2+ Signaling Requirements for Long-Term Depression in the Hippocampus. *Neuron*, 16:825–833.

Dai, J. and Wang, Y. (2018). Contrast coding in the primary visual cortex depends on temporal contexts. *Eur. J. Neurosci.*, 47(8):947–958.

Daw, N. W., Stein, P. S., and Fox, K. (1993). The role of NMDA receptors in information processing. *Annu. Rev. Neurosci.*, 16:207–22.

Dayan, P. and Abbott, L. F. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. The MIT Press.

De Roo, M., Klauser, P., and Muller, D. (2008). LTP promotes a selective long-term stabilization and clustering of dendritic spines. *PLoS Biol.*, 6(9):1850–1860.

De Valois, R. L., Albrecht, D. G., and Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Res.*, 22(5):545–59.

Deger, M., Helias, M., Rotter, S., and Diesmann, M. (2012). Spike-timing dependence of structural plasticity explains cooperative synapse formation in the neocortex. *PLoS Comput. Biol.*, 8(9):e1002689.

Deger, M., Seeholzer, A., and Gerstner, W. (2018). Multicontact Co-operativity in Spike-Timing–Dependent Structural Plasticity Stabilizes Networks. *Cereb. Cortex*, 28(4):1396–1415.

Desai, N. S., Cudmore, R. H., Nelson, S. B., and Turrigiano, G. G. (2002). Critical periods for experience-dependent synaptic scaling in visual cortex. *Nat. Neurosci.*, 5(8):783–9.

Desai, N. S., Rutherford, L. C., and Turrigiano, G. G. (1999). Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nat. Neurosci.*, 2(6):515–520.

DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.*, 11(8):333–41.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3):415–34.

Diehl, P. U. and Cook, M. (2015). Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Front. Comput. Neurosci.*, 9(August):1–9.

Douglas, R. J. and Martin, K. A. C. (2004). Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.*, 27:419–51.

Ecker, A. S., Berens, P., Keliris, G. a., Bethge, M., Logothetis, N. K., and Tolias, A. S. (2010). Decorrelated neuronal firing in cortical microcircuits. *Science*, 327(5965):584–7.

Einhäuser, W., Hipp, J., Eggert, J., Körner, E., and König, P. (2005). Learning viewpoint invariant object representations using a temporal coherence principle. *Biol. Cybern.*, 93(1):79–90.

Einhäuser, W., Kayser, C., König, P., and Körding, K. P. (2002). Learning the invariance properties of complex cells from their responses to natural stimuli. *Eur. J. Neurosci.*, 15(3):475–86.

Elliott, T. (2014). Sparseness, antisparseness and anything in between: the operating point of a neuron determines its computational repertoire. *Neural Comput.*, 26(9):1924–72.

Evans, B. D. and Stringer, S. M. (2012). Transformation-invariant visual representations in self-organizing spiking neural networks. *Front. Comput. Neurosci.*, 6(July):46.

Falconbridge, M. S., Stamps, R. L., and Badcock, D. R. (2006). A simple Hebbian/anti-Hebbian network learns the sparse, independent components of natural images. *Neural Comput.*, 18(2):415–29.

Fauth, M., Wörgötter, F., and Tetzlaff, C. (2015a). Formation and Maintenance of Robust Long-Term Information Storage in the Presence of Synaptic Turnover. *PLoS Comput. Biol.*, 11(12).

Fauth, M., Wörgötter, F., and Tetzlaff, C. (2015b). The Formation of Multi-synaptic Connections by the Interaction of Synaptic and Structural Plasticity and Their Functional Consequences. *PLoS Comput. Biol.*, 11(1):1–29.

Feldman, D. E. (2009). Synaptic mechanisms for plasticity in neocortex. *Annu. Rev. Neurosci.*, 32:33–55.

Felleman, D. J. and Van Essen, D. C. (1991). Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cereb. Cortex*, 1(1):1–47.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Comput.*, 6(4):559–601.

Földiák, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biol. Cybern.*, 237(5349):55–56.

Földiák, P. (1991). Learning Invariance from Transformation Sequences. *Neural Comput.*, 3(2):194–200.

Földiák, P. (1998). Learning constancies for object perception. In Walsh, V. and Kulikowski, J. J., editors, *Percept. Constancy Why things look as they do*, pages 144–172. Cambridge Univ. Press, Cambridge U.K.

Földiák, P. and Young, M. P. (1995). Sparse Coding in the Primate Cortex. In Arbib, M. A., editor, *Handb. brain theory neural networks*, pages 895–898. MIT Press, Cambridge, MA, USA.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., and Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nat. Neurosci.*, 16(7):974–81.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.*, 202(36):193 202.

Gallinaro, J. V. and Rotter, S. (2018). Associative properties of structural plasticity based on firing rate homeostasis in recurrent neuronal networks. *Sci. Rep.*, 8(1):1–13.

Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. (2009). *Cognitive neuroscience: the biology of the mind*. Norton.

Goold, C. P. and Nicoll, R. A. (2010). Single-cell optogenetic excitation drives homeostatic synaptic depression. *Neuron*, 68(3):512–28.

Graham, N. V. (2011). Beyond multiple pattern analyzers modeled as linear filters (as classical V1 simple cells): useful additions of the last 25 years. *Vision Res.*, 51(13):1397–430.

Harris, K. D. and Mrsic-Flogel, T. D. (2013). Cortical connectivity and sensory coding. *Nature*, 503(7474):51–58.

Harvey, C. D., Yasuda, R., Zhong, H., and Svoboda, K. (2008). The Spread of Ras Activity Triggered by Activation of a Single Dendritic Spine. *Science*, 321(5885):136–140.

Hashimoto, W. (2003). Quadratic forms in natural images. *Network*, 14(4):765–88.

Hebb, D. (1949). The Organization of Behavior. A Neuropsychological Theory. 1949. *Wiley, New York*.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Vis. Neurosci.*, 9(2):181–97.

Hegdé, J. and Van Essen, D. C. (2000). Selectivity for complex shapes in primate visual area V2. *J. Neurosci.*, 20(5):RC61.

Hegdé, J. and Van Essen, D. C. (2003). Strategies of shape representation in macaque visual area V2. *Vis. Neurosci.*, 20(3):313–28.

Helias, M. (2008). Structural plasticity controlled by calcium based correlation detection. *Front. Comput. Neurosci.*, 2(December):7.

Hirsch, J. A., Martinez, L. M., Pillai, C., Alonso, J.-M., Wang, Q., and Sommer, F. T. (2003). Functionally distinct inhibitory neurons at the first stage of visual cortical processing. *Nat. Neurosci.*, 6(12):1300–1308.

Hofer, S. B., Ko, H., Pichler, B., Vogelstein, J., Ros, H., Zeng, H., Lein, E., Lesica, N. A., and Mrsic-Flogel, T. D. (2011). Differential connectivity and response dynamics of excitatory and inhibitory neurons in visual cortex. *Nat. Neurosci.*, 14(8):1045–1052.

Holtmaat, A. J. G. D. and Svoboda, K. (2009). Experience-dependent structural synaptic plasticity in the mammalian brain. *Nat. Rev. Neurosci.*, 10(10):759–759.

Holtmaat, A. J. G. D., Trachtenberg, J. T., Wilbrecht, L., Shepherd, G. M., Zhang, X., Knott, G. W., and Svoboda, K. (2005). Transient and persistent dendritic spines in the neocortex in vivo. *Neuron*, 45(2):279–291.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, 5:1457–1469.

Hu, H., Shao, L. R., Chavoshy, S., Gu, N., Trieb, M., Behrens, R., Laake, P., Pongs, O., Knaus, H. G., Ottersen, O. P., and Storm, J. F. (2001). Presynaptic Ca2+-activated K+ channels in glutamatergic hippocampal terminals and their role in spike repolarization and regulation of transmitter release. *J. Neurosci.*, 21(24):9585–97.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, 160(1):106–154.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.*, 195(1):215.

Hyvärinen, A. and Hoyer, P. O. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Res.*, 41(18):2413–23.

Hyvärinen, a. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–30.

Intrator, N. and Cooper, L. N. (1992). Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5(1):3–17.

Isaacson, J. S. and Scanziani, M. (2011). How inhibition shapes cortical activity. *Neuron*, 72(2):231–243.

Ito, M. and Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *J. Neurosci.*, 24(13):3313–24.

Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.*, 58(6):1233–58.

Joshi, P. and Triesch, J. (2009). Rules for Information Maximization in Spiking Neurons Using Intrinsic Plasticity. *Neural Networks, 2009. IJCNN 2009.*, 8:1456–1461.

Kachergis, G., Wyatte, D., O'Reilly, R. C., de Kleijn, R., and Hommel, B. (2014). A continuous-time neural model for sequential action. *Philos. Trans. R. Soc. B Biol. Sci.*, 369(1655):20130623–20130623.

Kanan, C. (2007). NIMBLER: A Model of Visual Attention and Object Recognition With a Biologically Plausible Retina. *CSE 252C Fall 2007*, pages 1–8.

Karklin, Y. and Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457(7225):83–6.

Kasai, H., Fukuda, M., Watanabe, S., Hayashi-Takagi, A., and Noguchi, J. (2010). Structural dynamics of dendritic spines in memory and cognition. *Trends Neurosci.*, 33(3):121–9.

Katzner, S., Busse, L., and Carandini, M. (2011). GABAA inhibition controls response gain in visual cortex. *J. Neurosci.*, 31(16):5931–5941.

Kayser, C., Einhäuser, W., Dümmer, O., König, P., and Körding, K. P. (2001). Extracting slow subspaces from natural videos leads to complex cells. *Int. Conf. Artif. neural networks*, pages 1075–1080.

Kermani Kolankeh, A., Teichmann, M., and Hamker, F. H. (2015). Competition improves robustness against loss of information. *Front. Comput. Neurosci.*, 9(March):1–12.

Kheradpisheh, S. R., Ganjtabesh, M., and Masquelier, T. (2016). Bio-inspired unsupervised learning of visual features leads to robust invariant object recognition. *Neurocomputing*, 205:382–392.

Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J., and Masquelier, T. (2018). STDP-based spiking deep convolutional neural networks for object recognition. *Neural Networks*, 99:56–67.

King, P. D., Zylberberg, J., and DeWeese, M. R. (2013). Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1. *J. Neurosci.*, 33(13):5475–85.

Knott, G. W., Holtmaat, A. J. G. D., Wilbrecht, L., Welker, E., and Svoboda, K. (2006). Spine growth precedes synapse formation in the adult neocortex in vivo. *Nat. Neurosci.*, 9(9):1117–1124.

Kobatake, E. and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.*, 71(3):856–67.

Körding, K. P., Kayser, C., Einhäuser, W., and König, P. (2004). How are complex cell properties adapted to the statistics of natural stimuli? *J. Neurophysiol.*, 91(1):206–12.

Köster, U. and Hyvärinen, A. (2010). A two-layer model of natural stimuli estimated with score matching. *Neural Comput.*, 22(9):2308–33.

Kouh, M. and Poggio, T. (2008). A canonical neural circuit for cortical nonlinear operations. *Neural Comput.*, 20(6):1427–1451.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.*, pages 1–9.

Law, C. C. and Cooper, L. N. (1994). Formation of receptive fields in realistic visual environments according to the Bienenstock, Cooper, and Munro (BCM) theory. *Proc. Natl. Acad. Sci.*, 91(16):7797–801.

Le, Q. V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J., and Ng, A. Y. (2012). Building high-level features using large scale unsupervised learning. In *Proc. 29th Int. Coference Int. Conf. Mach. Learn.*, volume ICML'12, pages 507–514, Edinburgh, Scotland. Omnipress.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324.

Lee, H. and Ng, A. Y. (2008). Sparse deep belief net model for visual area V2. In *Adv. Neural Inf. Process. Syst. 20*, pages 873–880.

Lehky, S. R., Sejnowski, T. J., and Desimone, R. (2005). Selectivity and sparseness in the responses of striate complex cells. *Vision Res.*, 45(1):57–73.

Levy, W. B. and Baxter, R. A. (1996). Energy efficient neural codes. *Neural Comput.*, 8(3):531–43.

Lisman, J. (1989). A mechanism for the Hebb and the anti-Hebb processes underlying learning and memory. *Proc. Natl. Acad. Sci.*, 86(23):9574–8.

Lledo, P., Hjelmstad, G., Mukherji, S., Soderling, T., Malenka, R. C., and Nicoll, R. A. (1995). Calcium/calmodulin-dependent kinase II and long-term potentiation enhance

synaptic transmission by the same mechanism. *Proc. Natl. Acad. Sci.*, 92(24):11175–11179.

Malenka, R. C. and Bear, M. F. (2004). LTP and LTD: an embarrassment of riches. *Neuron*, 44(1):5–21.

Malinow, R., Schulman, H., and Tsien, R. W. (1989). Inhibition of postsynaptic PKC or CaMKII blocks induction but not expression of LTP. *Science*, 245(4920):862–6.

Martinez, L. M., Wang, Q., Reid, R. C., Pillai, C., Alonso, J.-M., Sommer, F. T., and Hirsch, J. A. (2005). Receptive field structure varies with layer in the primary visual cortex. *Nat. Neurosci.*, 8(3):372–9.

Martinez-Conde, S., Macknik, S. L., and Hubel, D. H. (2004). The role of fixational eye movements in visual perception. *Nat. Rev. Neurosci.*, 5(3):229–40.

Masquelier, T. (2012). Relative spike time coding and STDP-based orientation selectivity in the early visual system in natural continuous and saccadic vision: A computational model. *J. Comput. Neurosci.*, 32(3):425–441.

Masquelier, T. and Thorpe, S. J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput. Biol.*, 3(2):e31.

Matsuzaki, M., Honkura, N., Ellis-Davies, G. C. R., and Kasai, H. (2004). Structural basis of long-term potentiation in single dendritic spines. *Nature*, 429(6993):761–6.

Miconi, T., McKinstry, J. L., and Edelman, G. M. (2016). Spontaneous emergence of fast attractor dynamics in a model of developing primary visual cortex. *Nat Commun*, 7:13208.

Mobahi, H., Collobert, R., and Weston, J. (2009). Deep Learning from Temporal Coherence in Video. *26th Annu. Int. Conf. Mach. Learn.*, pages 737–744.

Morrison, A., Diesmann, M., and Gerstner, W. (2008). Phenomenological models of synaptic plasticity based on spike timing. *Biol. Cybern.*, 98(6):459–478.

Mutch, J. and Lowe, D. G. (2008). Object Class Recognition and Localization Using Sparse Features with Limited Receptive Fields. *Int. J. Comput. Vis.*, 80(1):45–57.

Nelson, S. B. and Turrigiano, G. G. (2008). Strength through diversity. *Neuron*, 60(3):477–82.

Nene, S., Nayar, S., and Murase, H. (1996). Columbia Object Image Library (COIL-100). Technical report, Columbia University, New York.

Neumann, K., Strub, C., and Steil, J. (2013). Intrinsic plasticity via natural gradient descent with application to drift compensation. *Neurocomputing*, 112:26–33.

Ohzawa, I. and Freeman, R. D. (1988). Binocularly deprived cats: binocular tests of cortical cells show regular patterns of interaction. *J. Neurosci.*, 8(7):2507–16.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J. Math. Biol.*, 15(3):267–273.

Olshausen, B. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–9.

Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res.*, 37(23):3311–25.

Olshausen, B. A. and Field, D. J. (2004). Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.*, 14(4):481–7.

Olshausen, B. A. and Lewicki, M. (2013). What Natural Scene Statistics Can Tell Us about Cortical Representation. In Werner, J. and Chalupa, L., editors, *New Vis. Neurosci.* MIT Press.

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., and Jilk, D. J. (2013). Recurrent processing during object recognition. *Front. Psychol.*, 4(April):124.

O'Reilly, R. C., Wyatte, D., and Rohrlich, J. (2014). Learning Through Time in the Thalamocortical Loops. *Arxiv:1407.3432 [Q-Bio]*.

Osindero, S., Welling, M., and Hinton, G. E. (2006). Topographic product models applied to natural scene statistics. *Neural Comput.*, 18(2):381–414.

Perry, G., Rolls, E. T., and Stringer, S. M. (2010). Continuous transformation learning of translation invariant representations. *Exp. Brain Res.*, 204(2):255–270.

Pettit, D., Perlman, S., and Malinow, R. (1994). Potentiated transmission and prevention of further LTP by increased CaMKII activity in postsynaptic hippocampal slice neurons. *Science*, 266(5192):1881–1885.

Portilla, J. and Simoncelli, E. P. (2000). A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficient. *Int. J. Comput. Vis.*, 40(1):49–71.

Potjans, T. C. and Diesmann, M. (2014). The cell-type specific cortical microcircuit: Relating structure and activity in a full-scale spiking network model. *Cereb. Cortex*, 24(3):785–806.

Rehn, M. and Sommer, F. T. (2007). A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *J. Comput. Neurosci.*, 22(2):135–46.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2(11):1019–25.

Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J. Neurophysiol.*, 88(1):455–463.

Ringach, D. L. and Malone, B. J. (2007). The operating point of the cortex: neurons as large deviation detectors. *J. Neurosci.*, 27(29):7673–83.

Ringach, D. L. and Shapley, R. (2004). Reverse correlation in neurophysiology. *Cogn. Sci.*, 28(2):147–166.

Ringach, D. L., Shapley, R. M., and Hawken, M. J. (2002). Orientation selectivity in macaque V1: diversity and laminar dependence. *J. Neurosci.*, 22(13):5639–5651.

Rockland, K. S. and Virga, A. (1989). Terminal arbors of individual "Feedback" axons projecting from area V2 to V1 in the macaque monkey: A study using immunohistochemistry of anterogradely transported Phaseolus vulgaris-leucoagglutinin. *J. Comp. Neurol.*, 285(1):54–72.

Rolfs, M. (2009). Microsaccades: Small steps on a long way. *Vision Res.*, 49(20):2415–2441.

Rolls, E. T. (2012). Invariant Visual Object and Face Recognition: Neural and Computational Bases, and a Model, VisNet. *Front. Comput. Neurosci.*, 6(June):35.

Rolls, E. T. and Milward, T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Comput.*, 12(11):2547–72.

Rolls, E. T. and Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J. Neurophysiol.*, 73(2):713–26.

Rolls, E. T. and Treves, A. (2011). The neuronal encoding of information in the brain. *Prog. Neurobiol.*, 95(3):448–490.

Rolls, E. T., Treves, A., Tovee, M. J., and Panzeri, S. (1997). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *J. Comput. Neurosci.*, 4(4):309–333.

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., and Olshausen, B. A. (2008). Sparse Coding via Thresholding and Local Competition in Neural Circuits. *Neural Comput.*, 20(10):2526–2563.

Ruthazer, E. S. and Aizenman, C. D. (2010). Learning to see: patterned visual activity and the development of visual function. *Trends Neurosci.*, 33(4):183–92.

Sadeh, S., Clopath, C., and Rotter, S. (2015). Emergence of Functional Specificity in Balanced Networks with Synaptic Plasticity. *PLOS Comput. Biol.*, 11(6):e1004307.

Sala, C. and Segal, M. (2014). Dendritic Spines: The Locus of Structural and Functional Plasticity. *Physiol. Rev.*, 94(1):141–188.

Savin, C., Joshi, P., and Triesch, J. (2010). Independent component analysis in spiking neurons. *PLoS Comput. Biol.*, 6(4):e1000757.

Schiller, P. H., Finlay, B. L., and Volman, S. F. (1976). Quantitative studies of single-cell properties in monkey striate cortex. I. Spatiotemporal organization of receptive fields. *J. Neurophysiol.*, 39(6):1288–319.

Sejnowski, T. J. (1977). Storing covariance with nonlinearly interacting neurons. *J. Math. Biol.*, 4(4):303–21.

Serre, T. (2006). *Learning a Dictionary of Shape-Components in Visual Cortex: Comparison with Neurons, Humans and Machines*. Doctor thesis, Massachusetts Institute of Technology.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):411–26.

Sharpee, T. O. (2013). Computational identification of receptive fields. *Annu. Rev. Neurosci.*, 36:103–20.

Shipp, S. (2003). The functional logic of cortico-pulvinar connections. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 358(1438):1605–24.

Shipp, S. (2007). Structure and function of the cerebral cortex. *Curr. Biol.*, 17(12):443–449.

Shipp, S. (2015). Pulvinar Structure, Circuitry & Function in Primates. In *Ref. Modul. Biomed. Sci.*, pages 1–27. Elsevier.

Shouval, H. Z., Bear, M. F., and Cooper, L. N. (2002a). A unified model of NMDA receptor-dependent bidirectional synaptic plasticity. *Proc. Natl. Acad. Sci.*, 99(16):10831.

Shouval, H. Z., Castellani, G. C., Blais, B. S., Yeung, L. C., and Cooper, L. N. (2002b). Converging evidence for a simplified biophysical model of synaptic plasticity. *Biol. Cybern.*, 87(5-6):383–91.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annu. Rev. Neurosci.*, 24(1):1193–1216.

Sincich, L. C. and Horton, J. C. (2005). The Circuitry of V1 and V2: Integration of Color, Form, and Motion. *Annu. Rev. Neurosci.*, 28(1):303–326.

Sjöström, P. J., Turrigiano, G. G., and Nelson, S. B. (2001). Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron*, 32(6):1149–64.

Smith, M. A. and Kohn, A. (2008). Spatial and Temporal Scales of Neuronal Correlation in Primary Visual Cortex. *J. Neurosci.*, 28(48):12591–12603.

Spanne, A. and Jörntell, H. (2015). Questioning the role of sparse coding in the brain. *Trends Neurosci.*, 38(7):417–427.

Spratling, M. W. (2005). Learning viewpoint invariant perceptual representations from cluttered images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(5):753–61.

Spratling, M. W. (2012). Unsupervised Learning of Generative and Discriminative Weights Encoding Elementary Image Components in a Predictive Coding Model of Cortical Function. *Neural Comput.*, 24(1):60–103.

Spratling, M. W. (2017). A Hierarchical Predictive Coding Model of Object Recognition in Natural Images. *Cognit. Comput.*, 9(2):151–167.

Stemmler, M. and Koch, C. (1999). How voltage-dependent conductances can adapt to maximize the information encoded by neuronal firing rate. *Nat. Neurosci.*, 2(6):521–527.

Stepanyants, A. and Chklovskii, D. B. (2005). Neurogeometry and potential synaptic connectivity. *Trends Neurosci.*, 28(7):387–394.

Stepanyants, A., Hof, P. R., and Chklovskii, D. B. (2002). Geometry and structural plasticity of synaptic connectivity. *Neuron*, 34(2):275–88.

Stringer, S. M., Perry, G., Rolls, E. T., and Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biol. Cybern.*, 94(2):128–42.

Stringer, S. M. and Rolls, E. T. (2008). Learning transform invariant object recognition in the visual system with multiple stimuli present during training. *Neural Networks*, 21(7):888–903.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going Deeper with Convolutions. *CoRR*, 07-June-20:1–9.

Teichmann, M. and Hamker, F. H. (2015). Intrinsic Plasticity: A Simple Mechanism to Stabilize Hebbian Learning in Multilayer Neural Networks. In Villmann, T. and Schleif, F.-M., editors, *Mach. Learn. Reports 03/2015*, pages 103–111.

Teichmann, M. and Hamker, F. H. (2017). Learning Stable Recurrent Excitation in Simulated Biological Neural Networks. In Lintas, A., Rovetta, S., Verschure, P. F., and Villa, A. E., editors, *Artif. Neural Networks Mach. Learn. – ICANN 2017*, volume 10613 of *Lecture Notes in Computer Science*, pages 449–450, Cham. Springer International Publishing.

Teichmann, M., Wiltschut, J., and Hamker, F. H. (2012). Learning invariance from natural images inspired by observations in the primary visual cortex. *Neural Comput.*, 24(5):1271–96.

Thomson, A. M. (2010). Neocortical layer 6, a review. *Front. Neuroanat.*, 4:1–14.

Thomson, A. M. and Bannister, A. P. (2003). Interlaminar connections in the neocortex. *Cereb. Cortex*, 13(1):5–14.

Tolhurst, D. J., Smyth, D., and Thompson, I. D. (2009). The Sparseness of Neuronal Responses in Ferret Primary Visual Cortex. *J. Neurosci.*, 29(8):2355–2370.

Treves, A. and Rolls, E. T. (1991). What determines the capacity of autoassociative memories in the brain? *Netw. Comput. Neural Syst.*, 2(4):371–397.

Triesch, J. (2005a). A Gradient Rule for the Plasticity of a Neuron's Intrinsic Excitability. In Duch, W., Oja, E., and Zadrozny, S., editors, *Int. Conf. Artif. Neural Networks*, volume 3696 LNCS, pages 65–70. Springer-Verlag Berlin Heidelberg, Warsaw, Poland.

Triesch, J. (2005b). Synergies between Intrinsic and Synaptic Plasticity in Individual Model Neurons. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Adv. Neural Inf. Process. Syst. 17*, pages 1417–1424. MIT Press.

Triesch, J. (2007). Synergies between intrinsic and synaptic plasticity mechanisms. *Neural Comput.*, 19(4):885–909.

Turrigiano, G. G. (2011). Too many cooks? Intrinsic and synaptic homeostatic mechanisms in cortical circuit refinement. *Annu. Rev. Neurosci.*, 34:89–103.

Turrigiano, G. G., Leslie, K. R., Desai, N. S., Rutherford, L. C., and Nelson, S. B. (1998). Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*, 391(6670):892–6.

Turrigiano, G. G. and Nelson, S. B. (2004). Homeostatic plasticity in the developing nervous system. *Nat. Rev. Neurosci.*, 5(2):97–107.

van Hateren, J. H. and Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc. R. Soc. B Biol. Sci.*, 265(1):2315–20.

van Hateren, J. H. and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Biol. Sci.*, 265(1394):359–66.

Vinje, W. E. (2000). Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision. *Science*, 287(5456):1273–1276.

Vislay-Meltzer, R. L., Kampff, A. R., and Engert, F. (2006). Spatiotemporal Specificity of Neuronal Activity Directs the Modification of Receptive Fields in the Developing Retinotectal System. *Neuron*, 50(1):101–114.

Vitay, J., Dinkelbach, H. Ü., and Hamker, F. H. (2015). ANNarchy: a code generation approach to neural simulations on parallel hardware. *Front. Neuroinform.*, 9(July):19.

Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C., and Gerstner, W. (2011). Inhibitory Plasticity Balances Excitation and Inhibition in Sensory Pathways and Memory Networks. *Science*, 334(6062):1569–1573.

Wallis, G. and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Prog. Neurobiol.*, 51(2):167–94.

Wandell, B. A. and Winawer, J. (2015). Computational neuroimaging and population receptive fields. *Trends Cogn. Sci.*, 19(6):349–357.

Weber, C. and Triesch, J. (2008). A sparse generative model of V1 simple cells with intrinsic plasticity. *Neural Comput.*, 20(5):1261–84.

Wersing, H. and Koerner, E. (2003). Learning Optimized Features for Hierarchical Modelling of Invariant Object Recognition. *Neural Comput.*, 15(7):1–27.

Willmore, B. D. B., Bulstrode, H., and Tolhurst, D. J. (2012). Contrast normalization contributes to a biologically-plausible model of receptive-field development in primary visual cortex (V1). *Vision Res.*, 54:49–60.

Willmore, B. D. B., Mazer, J. A., and Gallant, J. L. (2011). Sparse coding in striate and extrastriate visual cortex. *J. Neurophysiol.*, 105(6):2907–2919.

Willmore, B. D. B., Prenger, R. J., and Gallant, J. L. (2010). Neural representation of natural images in visual area V2. *J. Neurosci.*, 30(6):2102–2114.

Willmore, B. D. B. and Tolhurst, D. J. (2001). Characterizing the sparseness of neural codes. *Netw. Comput. Neural Syst.*, 12(3):255–270.

Wiltschut, J. and Hamker, F. H. (2009). Efficient coding correlates with spatial frequency tuning in a model of V1 receptive field organization. *Vis. Neurosci.*, 26(1):21–34.

Xing, D., Ringach, D. L., Hawken, M. J., and Shapley, R. M. (2011). Untuned suppression makes a major contribution to the enhancement of orientation selectivity in macaque v1. *J. Neurosci.*, 31(44):15972–82.

Yang, S.-N., Tang, Y.-G., and Zucker, R. S. (1999). Selective Induction of LTP and LTD by Postsynaptic [Ca2+] Elevation. *J. Neurophysiol.*, 81:781–787.

Yasumatsu, N., Matsuzaki, M., Miyazaki, T., Noguchi, J., and Kasai, H. (2008). Principles of long-term dynamics of dendritic spines. *J. Neurosci.*, 28(50):13592–608.

Yeung, L. C., Shouval, H. Z., Blais, B. S., and Cooper, L. N. (2004). Synaptic homeostasis and input selectivity follow from a calcium-dependent plasticity model. *Proc. Natl. Acad. Sci.*, 101(41):14943–8.

Yoshimura, Y. and Callaway, E. M. (2005). Fine-scale specificity of cortical networks depends on inhibitory cell type and connectivity. *Nat. Neurosci.*, 8(11):1552–1559.

Zenke, F. and Gerstner, W. (2017). Hebbian plasticity requires compensatory processes on multiple timescales. *Philos. Trans. R. Soc. B Biol. Sci.*, 372(1715):20160259.

Zenke, F., Gerstner, W., and Ganguli, S. (2017). The temporal paradox of Hebbian learning and homeostatic plasticity. *Curr. Opin. Neurobiol.*, 43(April):166–176.

Zhang, W. and Linden, D. J. (2003). The other side of the engram: experience-driven changes in neuronal intrinsic excitability. *Nat. Rev. Neurosci.*, 4(11):885–900.

Zheng, P., Dimitrakakis, C., and Triesch, J. (2013). Network Self-Organization Explains the Statistics and Dynamics of Synaptic Connection Strengths in Cortex. *PLoS Comput. Biol.*, 9(1):e1002848.

Zou, W. Y., Zhu, S., NG, A. Y., and Yu, K. (2012). Deep Learning of Invariant Features via Simulated Fixations in Video. *Neural Inf. Process. Syst.*, pages 1–9.

Zylberberg, J., Murphy, J. T., and DeWeese, M. R. (2011). A Sparse Coding Model with Synaptically Local Plasticity and Spiking Neurons Can Account for the Diverse Shapes of V1 Simple Cell Receptive Fields. *PLoS Comput. Biol.*, 7(10):e1002250.

Appendix A

Model Parameter

A.1 Activity dependent spatial growth model

Parameter	Meaning	Value
c_s	scaling constant	0.01
d	L^∞ distance defining the size of the neighborhood	3
t^{basis}	basis interval	1000ms
Δt	time since last update	20000ms

TABLE A.1: Parameters for synapse creation.

Parameter	Meaning	Value
$p^{del_{max}}$	maximal deletion probability	0.005
a	controls the slope	0.1
t^{basis}	basis interval	1000ms
Δt	time since last update	20000ms
	Excitatory synapses	
w^{half}	weight value with half the maximum deletion probability	0.01
	Inhibitory synapses	
w^{half}	weight value with half the maximum deletion probability	0.15

TABLE A.2: Parameters for synapse removal.

A.2 Neuron model with intrinsic parameters

Parameter	Meaning	Value / Initial value
τ_m	time constant of the membrane potential	10ms
m	membrane potential	0
r	firing rate	0
θ	threshold	0
a	slope	1

TABLE A.3: Parameters of the activation function. For all neurons.

A.3 Intrinsic plasticity mechanisms

Parameter	Meaning	Value
ε	drift value	0.01
θ_{target}	target value for the threshold	population mean over r
τ_θ	time constant of the threshold	10000ms
a_{target}	target value for the slope	population mean over r^2
τ_a	time constant of the slope	10000ms

TABLE A.4: Parameters of the intrinsic plasticity. For all neurons, except LGN.

A.4 Synaptic plasticity and homeostatic regulations

Parameter	Meaning	Value
$p^{del_{max}}$	maximal deletion probability	0.005
a	controls the slope	0.1
t^{basis}	basis interval	1000ms
Δt	time since last update	20000ms
	Layers 4 and inhibitory neurons	
τ_{Ca}	time constant of the calcium level	10ms
	Layers 2/3	
τ_{Ca}	time constant of the calcium level	500ms

TABLE A.5: Parameters of the neuronal calcium level.

Parameter	Meaning	Value
a	base value	5000
b	adder	30000
c	slope	10

TABLE A.6: Parameters for the time constant for calcium dependent synaptic change. For all neurons, except LGN.

Parameter	Meaning	Value
τ_α	time constant	10000ms
ε	small constant decay	0.0005
α_θ	excitation threshold	2
α_j	initial value	2
τ_H	time constant	100ms
$K = 0.05$	small constant decay	0.05
$\gamma = 0.7$	activity threshold	0.7
H_j	initial value	0

TABLE A.7: Parameters for the homeostatic regulation. For all neurons, except LGN.

Parameter	Meaning	Value
τ_c	time constant	4000
α_c	constant	0.1
θ_c	small threshold	0.05

TABLE A.8: Parameters for the anti-Hebbian plasticity. For all neurons, except LGN.

Appendix B

Methods

B.1 Visualization of V1-L4 receptive fields

We use the weight matrix from LGN to the regarded V1-L4 neuron or the matrix obtained with reverse correlation. The matrices consist of two 2D-planes, one with the weights from the on-center LGN neurons and one with the weights from the off-center LGN neurons. To obtain a matrix for visualization or fitting, we subtract both planes from each other, i.e. on-plane minus off-plane (cf. Wiltschut and Hamker, 2009). This is possible because the subfields have nearly no overlap.

We visualize the matrices using different normalizations or color maps. A gray scaled color map (cf. Wiltschut and Hamker, 2009; Rehn and Sommer, 2007) or a red-blue color map (cf. Cossell et al., 2015). The gray scaled color map shows bright values for strong weights from on-center LGN neurons and dark values for strong weights from off-center LGN neurons. Zero or non existing weights have a medium gray tone. The color map is symmetric, i.e. when the weights from one plane are weaker than they will have lower contrast. Similarly, the red-blue color map shows a bold red for strong weights from on-center LGN neurons and a bold blue for strong weights from off-center LGN neurons. Also this map is symmetric, here zero or non existing weights will be white. For better visibility of the the weights in the both planes, we normalize the weights from both planes individually (will be mentioned in the figure description), so that the full contrast is used. To visualize the weights of multiple neurons, we generate a figure with tiles, separated by small borders. Each tile represents the weights of one neuron.

B.2 Gabor fit

A Gabor function is defined as a product of a cosine with frequency f and phase ψ, weighted with a Gaussian with the extents σ_x and σ_y and the orientation θ. Additionally, an amplitude A to capture the signal range and an offset B to account for shifts in the baseline are added.

$$g(x,y,\sigma_x,\sigma_y,f,\theta,\psi,A,B) = A \cdot e^{\left(-\frac{x'^2}{2\sigma_x^2}-\frac{y'^2}{2\sigma_y^2}\right)} \cdot \cos(2\pi f x' - \psi) + B$$
$$x' = x \cdot \cos(\theta) + y \cdot \sin(\theta)$$
$$y' = -x \cdot \sin(\theta) + y \cdot \cos(\theta)$$

We used the best fit obtained with the MATLAB function *lsqnonlin* from 100 randomly chosen starting points. As input to the function we used the feedforward weights from LGN to V1-layer 4. We subtract the weights from the off-center LGN neurons from the on-center neurons (cf. Wiltschut and Hamker, 2009).

B.3 Gauss fit

The Gauss function has the spatial extents σ_x and σ_y and the orientation θ. Additionally, an amplitude A to capture the signal range and an offset B to account for shifts in the baseline are added.

$$gauss(x,y,\sigma_x,\sigma_y,\theta,A,B) = A \cdot e^{\left(-\frac{x'^2}{2\sigma_x^2}-\frac{y'^2}{2\sigma_y^2}\right)} + B$$
$$x' = x \cdot \cos(\theta) + y \cdot \sin(\theta)$$
$$y' = -x \cdot \sin(\theta) + y \cdot \cos(\theta)$$

We used the best fit obtained with the MATLAB function *lsqnonlin* from 50 randomly chosen starting points. As input to the function we used the spatial response maps obtained by shifts of the optimal stimulus. The maximal shift distance is 12 pixel for all directions.

B.4 Weight vector length following Oja

Local Hebbian learning with weight normalization can be described after Oja (1982) as the following term (Eqn. B.1).

$$\tau \frac{d\mathbf{w}}{dt} = \mathbf{u}v - \alpha v^2 \mathbf{w} \tag{B.1}$$

The weight change is determined by the coactivity of presynaptic u and postsynaptic activity v (the Hebbian term) and a normalization term depending on the postsynaptic activity, the weight w and a constant α (the Oja normalization). The Oja normalization restricts the length of the weight vector (L2-norm) relative to α (Eqn. B.2) (Oja, 1982; Dayan and Abbott, 2001).

$$\tau \frac{d|\mathbf{w}|^2}{dt} = \frac{1}{\alpha} \tag{B.2}$$

This can be shown by expanding Eqn. B.1 by $2\mathbf{w}$ following Dayan and Abbott (2001, Section 8.2, p. 11).

$$\tau \frac{d|\mathbf{w}|^2}{dt} = 2v \cdot \mathbf{u}\mathbf{w} - 2\alpha v^2 |\mathbf{w}|^2) \tag{B.3}$$

So that we get,

$$\tau \frac{d|\mathbf{w}|^2}{dt} = 2v^2(1 - \alpha|\mathbf{w}|^2) \tag{B.4}$$

whereby the resulting term $\mathbf{uw}$ is equal to the postsynaptic activity v. Further it can be seen that the postsynaptic activity, which is positive defined, is just a factor for the amount of weight change given by the normalization. Now we set the left side to zero to obtain the equilibrium state and we get Eqn. B.2.

When having inhibition in the system the calculation appears similar. The postsynaptic activity v can be understood as $\mathbf{uw} - I_{Inh}$, where I_{Inh} is the inhibitory current a neuron receives. Thus, we get

$$\tau \frac{d|\mathbf{w}|^2}{dt} = 2(v + I_{Inh})v - 2\alpha v^2 |\mathbf{w}|^2) \tag{B.5}$$

which gives us

$$\tau \frac{d|\mathbf{w}|^2}{dt} = 2v^2(1 + \frac{I_{Inh}}{v} - \alpha|\mathbf{w}|^2)) \tag{B.6}$$

So the weight vector relaxes to

$$\tau \frac{d|\mathbf{w}|^2}{dt} = \frac{1}{\alpha} + \frac{I_{Inh}}{\alpha v} \tag{B.7}$$

We can do the very likely assumptions that the inhibition I_{Inh} is relative to the activity and inhibition is by average not larger than the excitation a neuron receives. Thus, we get something in the range of $[0,1]$ for $\frac{I_{Inh}}{v}$ in systems with inhibition. So that the length of the weight vector will relax to a value in the range $[\frac{1}{\alpha}, \frac{2}{\alpha}]$.

Appendix C

Intrinsic Plasticity

C.1 Development of threshold and slope during learning

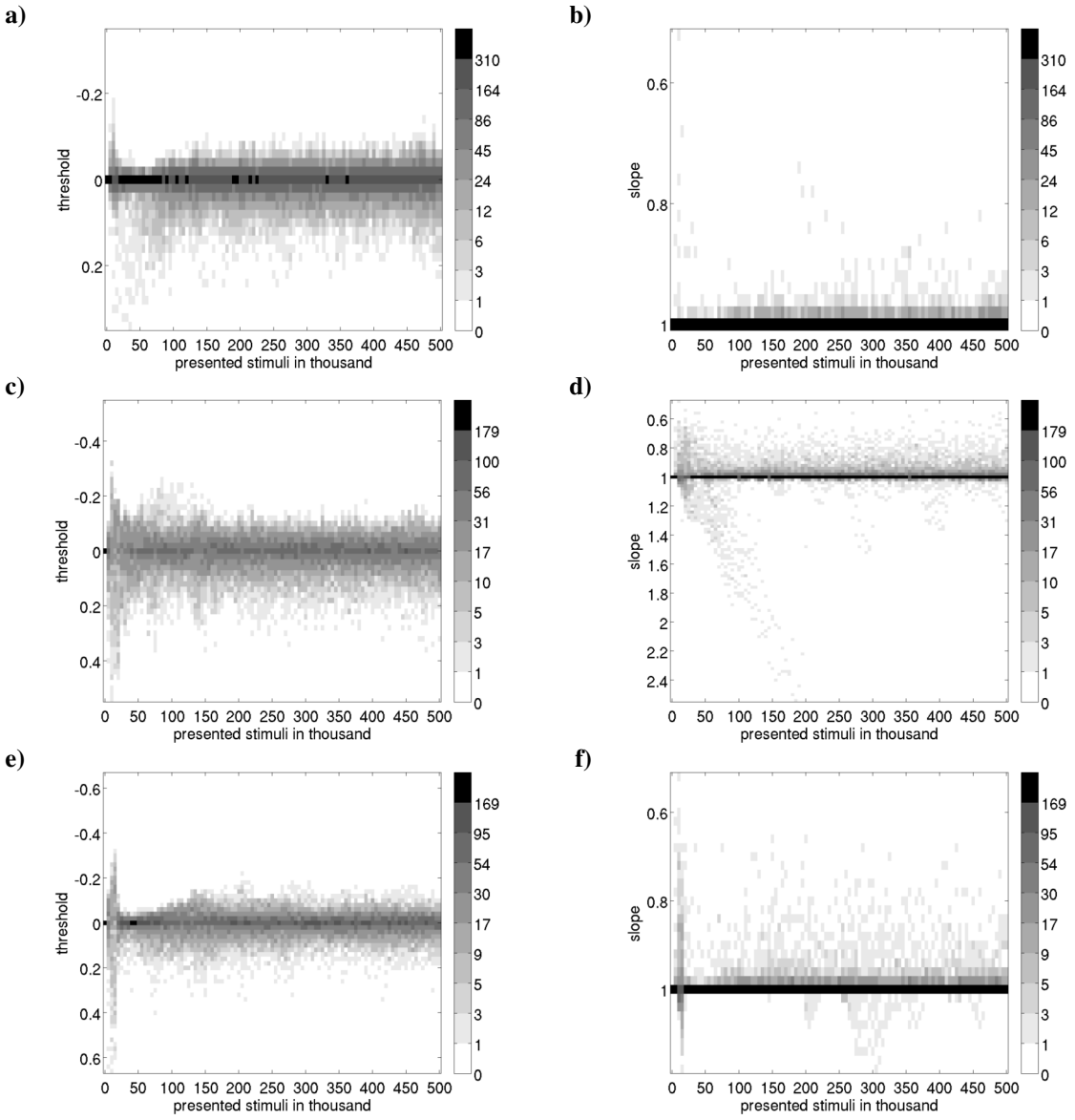

FIGURE C.1: Development of the intrinsic regulation parameters over time. The left column **ace)** shows the histogram of the threshold parameter θ for the excitatory neurons in the layers V1-L2/3, V2-L4, V2-L2/3 at each 5000 stimulus presentations over the full network training. The right column **bdf)** shows the slope parameter a of the same layers. In the early phases with strong synaptic plasticity the intrinsic regulation is strong. In later phases of the network development the values cluster closely around the parameter's origin. The gray tone indicate the amount of neurons within a bin. The bin size is 0.02. Note, the color scale is logarithmic to improve the visibility of the parameter distribution.

C.2 Histograms of the neuronal activity

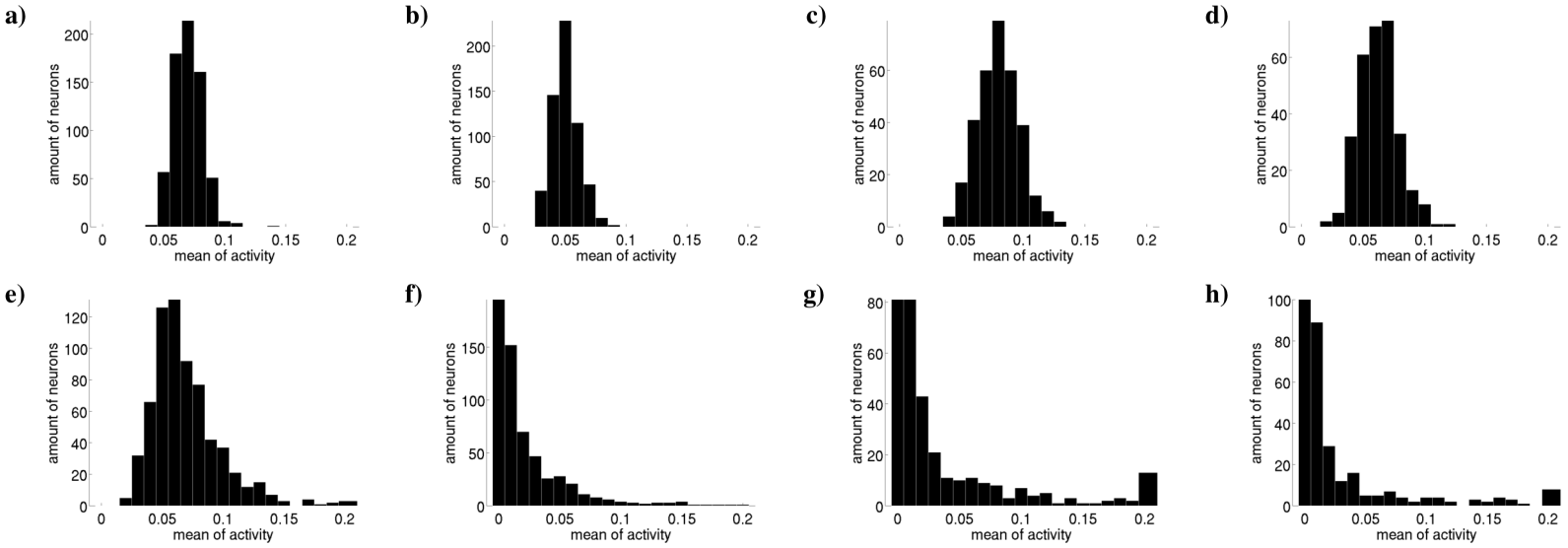

FIGURE C.2: Histograms of the mean of the neurons activity. The left column (a, c, e, g) shows results from layer 4 neurons and the right from layer 2/3 neurons. The mean of activity (a, b) and the variance of activity (e, f) are obtained with full intrinsic plasticity. Whereas the subsequent row (c, d; g, h) shows the results obtained without intrinsic plasticity.

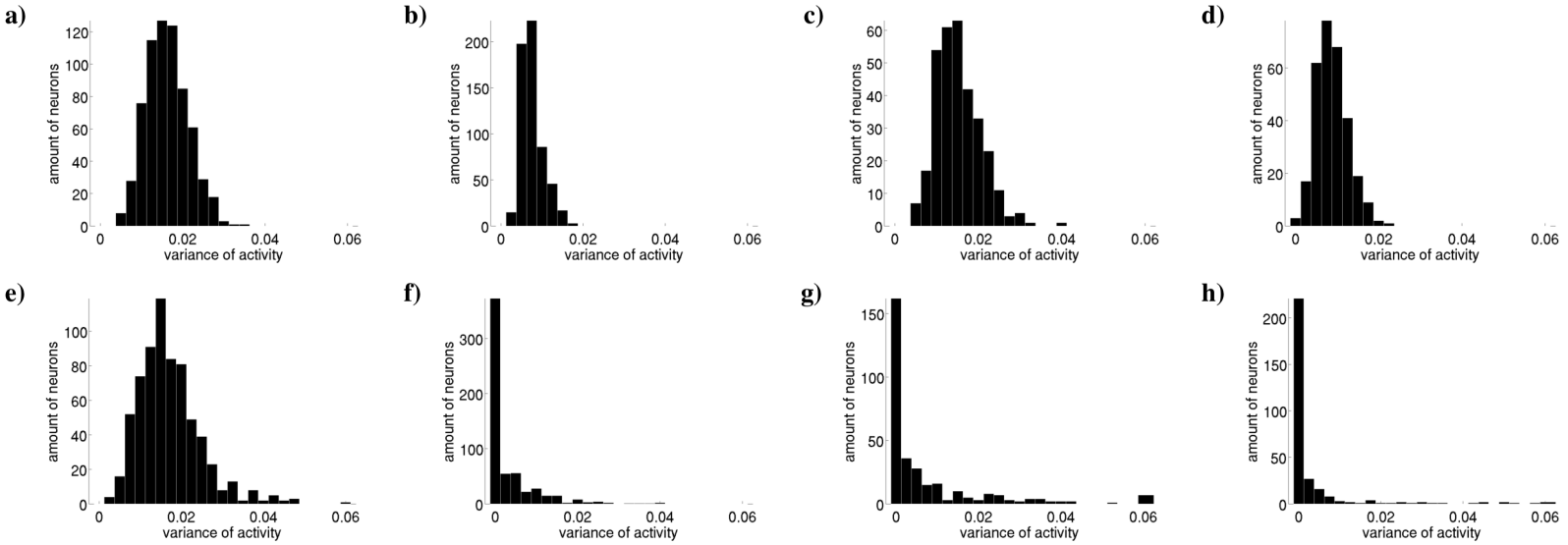

FIGURE C.3: Histograms of the variance of the neurons activity. The left column (a, c, e, g) shows results from layer 4 neurons and the right from layer 2/3 neurons. The mean of activity (a, b) and the variance of activity (e, f) are obtained with full intrinsic plasticity. Whereas the subsequent row (c, d; g, h) shows the results obtained without intrinsic plasticity.

Appendix D

Synaptic Plasticity

D.1 Efficient coding

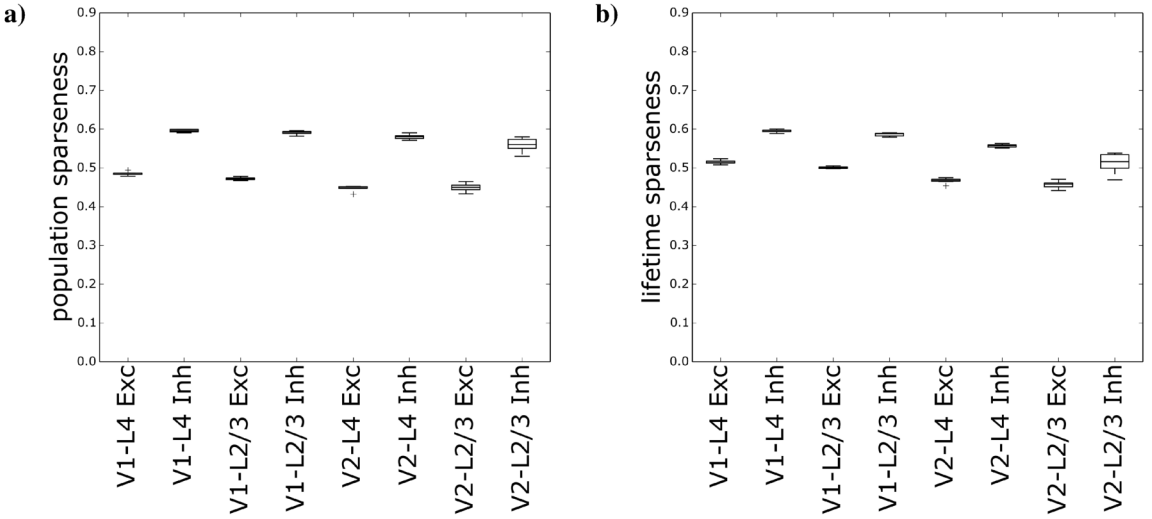

FIGURE D.1: Sparseness and correlations of the neurons in every population. The sparseness and correlations between the responses of all neurons on 100000 natural scene stimuli have been measured with the sparseness measure of Hoyer (2004) and linear correlations. a) Shows the box plot for the sparseness values each neuron population had on the stimuli, called population sparseness. b) The sparsenesses of the response of every single neuron in every population, called lifetime sparseness. c) The linear correlations of the responses between all pairs of neurons within a each population. The sparseness slightly decreases in deeper layers. Inhibitory populations develop higher sparseness than excitatory one. The correlations behave similarly, inhibitory populations show less correlations than excitatory populations. In general correlation values are very low, while the sparseness values are less sparse than assumed in sparse coding models.

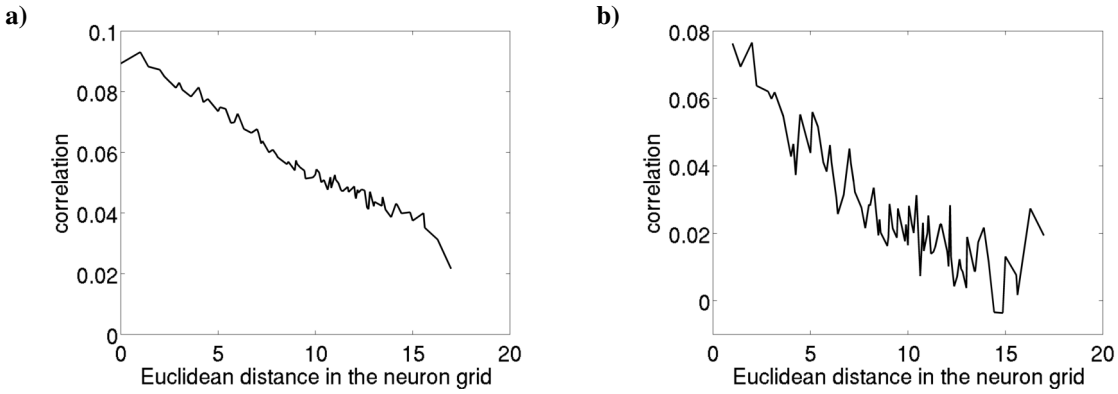

FIGURE D.2: Correlations of the neurons in relation to their distance in the neuron grid. a) Measured in the excitatory neurons of V1-L4 and **b)** in the inhibitory neurons.

D.2 Object recognition performance

Input	**10°**	**15°**	**30°**	**45°**	**60°**	**90°**
Colored	100.0	99.9167	98.4833	96.0312	91.8788	85.5441
Gray scaled	99.8889	98.9792	95.1167	90.6719	85.1364	75.2941
Gray scaled & whitened	96.4444	94.1250	87.0333	78.3906	71.1970	58.7206

TABLE D.1: Recognition accuracy on the COIL-100 dataset using different preprocessing steps. We trained a SVM with linear kernel with the raw and preprocessed images, using particular views from the dataset. We tested on all unused views. First we train with the original colored images. Second, with the gray scaled images. Third, with the whitened gray scaled and normalized images, which have been used as network input.

Layer	Type	Training views					
		10°	**15°**	**30°**	**45°**	**60°**	**90°**
V1-L4	Excitatory	98.2000 ± 0.1067	96.1833 ± 0.1823	90.1550 ± 0.2003	82.8469 ± 0.2663	74.5682 ± 0.3222	62.2676 ± 0.3080
	Inhibitory	97.9083 ± 0.0971	95.7604 ± 0.1254	88.2433 ± 0.2298	80.4375 ± 0.2628	71.9909 ± 0.3720	60.2559 ± 0.4723
	Exc & Inh	97.9139 ± 0.1028	95.7875 ± 0.1167	88.2850 ± 0.2403	80.4859 ± 0.2584	72.0364 ± 0.3842	60.2750 ± 0.4598
V1-L2/3	Excitatory	98.8028 ± 0.0685	97.2083 ± 0.1398	92.0033 ± 0.1584	85.4609 ± 0.2575	77.7182 ± 0.2828	65.4235 ± 0.4476
	Inhibitory	98.7778 ± 0.0994	97.0896 ± 0.1071	91.6667 ± 0.2047	84.9484 ± 0.2194	76.8333 ± 0.3964	64.3397 ± 0.4253
	Exc & Inh	98.7889 ± 0.1048	97.1083 ± 0.1028	91.7117 ± 0.1889	84.9891 ± 0.2220	76.8924 ± 0.3973	64.3765 ± 0.4117
V2-L4	Excitatory	98.9028 ± 0.0705	97.3479 ± 0.1476	92.4367 ± 0.1152	86.4453 ± 0.2274	78.5182 ± 0.3516	67.2441 ± 0.2880
	Inhibitory	98.7000 ± 0.1464	96.9646 ± 0.1141	91.2900 ± 0.2445	84.5641 ± 0.3441	76.5030 ± 0.4292	64.7588 ± 0.5151
	Exc & Inh	98.7000 ± 0.1442	96.9750 ± 0.1140	91.3333 ± 0.2153	84.6078 ± 0.3457	76.5500 ± 0.4228	64.7926 ± 0.5368
V2-L2/3	Excitatory	99.1139 ± 0.0809	97.6833 ± 0.1207	93.1567 ± 0.1689	87.8641 ± 0.3113	80.0439 ± 0.3783	69.1926 ± 0.4131
	Inhibitory	99.0083 ± 0.2236	97.4333 ± 0.3480	92.3500 ± 0.7738	86.6297 ± 1.0405	79.1803 ± 1.0899	67.9162 ± 0.8608
	Exc & Inh	99.0278 ± 0.2201	97.4479 ± 0.3610	92.4300 ± 0.7995	86.6937 ± 1.0645	79.2545 ± 1.0970	67.9544 ± 0.8525

TABLE D.2: Recognition accuracies and standard deviations on COIL-100 of all network populations. We trained a SVM with linear kernel on the images from the COIL-100 dataset, using particular views. We tested on all unused views. We show the mean accuracies and its standard deviations for each network population, obtained from 10 independent network runs.

Layer	Type	Training views					
		10°	**15°**	**30°**	**45°**	**60°**	**90°**
V1-L4	Excitatory	98.3796 ± 0.1370	96.5556 ± 0.0789	90.3667 ± 0.1424	83.5885 ± 0.0592	75.6768 ± 0.4980	63.3775 ± 0.4145
	Inhibitory	98.1306 ± 0.1309	96.0708 ± 0.1360	88.7367 ± 0.3032	80.6047 ± 0.3190	72.5242 ± 0.4509	61.2853 ± 0.3391
	Exc & Inh	98.1306 ± 0.1309	96.0708 ± 0.1360	88.7367 ± 0.3032	80.6047 ± 0.3190	72.5242 ± 0.4509	61.2853 ± 0.3393
V1-L2/3	Excitatory	98.3426 ± 0.0976	96.3611 ± 0.0434	89.5944 ± 0.2263	82.0104 ± 0.0393	73.2778 ± 0.3376	60.7745 ± 0.1620
	Inhibitory	98.8833 ± 0.0444	97.3604 ± 0.0946	92.2850 ± 0.2355	85.5281 ± 0.2590	77.2652 ± 0.4469	66.1029 ± 0.2948
	Exc & Inh	98.8833 ± 0.0444	97.3604 ± 0.0946	92.2850 ± 0.2355	85.5281 ± 0.2590	77.2652 ± 0.4469	66.1044 ± 0.2936
V2-L4	Excitatory	98.5185 ± 0.1052	96.8542 ± 0.0625	91.3667 ± 0.2466	85.3958 ± 0.2818	77.5505 ± 0.6972	67.0637 ± 0.7832
	Inhibitory	98.6667 ± 0.1052	96.9750 ± 0.1487	91.3917 ± 0.3249	84.1844 ± 0.3920	76.3712 ± 0.6282	65.4250 ± 0.5006
	Exc & Inh	98.6667 ± 0.1344	96.9750 ± 0.1487	91.3917 ± 0.3249	84.1844 ± 0.3920	76.3712 ± 0.6282	65.4235 ± 0.4990
V2-L2/3	Excitatory	98.3426 ± 0.1697	96.5000 ± 0.1458	90.1667 ± 0.4640	83.1094 ± 0.2817	74.6566 ± 0.5985	63.2402 ± 0.2488
	Inhibitory	98.6417 ± 0.3351	96.8937 ± 0.5616	91.0883 ± 1.4251	84.7734 ± 1.6483	77.2652 ± 1.5381	66.4838 ± 1.3187
	Exc & Inh	98.6417 ± 0.3351	96.8937 ± 0.5616	91.0883 ± 1.4251	84.7734 ± 1.6483	77.2636 ± 1.5368	66.4853 ± 1.3194

TABLE D.3: Recognition accuracies and standard deviations on COIL-100 of all network populations for the fast trace model. We trained a SVM with linear kernel on the images from the COIL-100 dataset, using particular views. We tested on all unused views. We show the mean accuracies and its standard deviations for each network population, obtained from three independent network runs. The model is similar configured to our standard model but uses a fast calcium trace ($\tau_{Ca} = 10$), so that the trace has no effect on learning.